DDSAnalytics

Anish Bhandari

2022-11-26

Introduction

DDSAnalytics is an analytics company that specializes in talent management solutions for Fortune 100 companies. In this document, we will analyze data and subsequently provide meaningful interpretations for our client Frito Lay. As a representative of DDSAnalytics, I'll meet with CEO and CFO to present my findings as well as recommendations on Dec 11,2022.

Data Collection

```
# Loading Data From S3 Objects Using the aws.s3 package

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.8     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.1
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(aws.s3)
```

```
## Warning: package 'aws.s3' was built under R version 4.2.2
```

```
library(readxl)

Sys.setenv("AWS_ACCESS_KEY_ID" = "AKIAVW4VHW4VPB7MHZWA",
           "AWS_SECRET_ACCESS_KEY" = "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",
           "AWS_DEFAULT_REGION" = "us-east-2")



# Using aws.s3
aws.s3::bucketlist()
```

```
##                Bucket             CreationDate
## 1    smuddsproject2 2022-07-26T14:50:49.000Z
## 2         smuds6306 2021-11-16T13:35:36.000Z
## 3 smuds6306breakout 2021-11-17T00:23:04.000Z
```

```
aws.s3::get_bucket("smuddsproject2")
```

```
## Bucket: smuddsproject2
##
## $Contents
## Key:            Case2PredictionsClassifyEXAMPLE.csv
## LastModified:   2022-07-18T19:01:53.000Z
## ETag:           "bd1de75effe9449f7d49a4de5116205a"
## Size (B):       3012
## Owner:          6a1a843f5cdc0cd887a7117c91d8bb3e0c9d31ba5db7441f7d01d6ade80cfac3
## Storage class:  STANDARD
##
## $Contents
## Key:            Case2PredictionsRegressEXAMPLE.csv
## LastModified:   2022-07-18T19:01:51.000Z
## ETag:           "a0f1f01c30e2cd00488822ad3c9aa6fe"
## Size (B):       3187
## Owner:          6a1a843f5cdc0cd887a7117c91d8bb3e0c9d31ba5db7441f7d01d6ade80cfac3
## Storage class:  STANDARD
##
## $Contents
## Key:            CaseStudy2-data.csv
## LastModified:   2022-07-18T19:00:38.000Z
## ETag:           "d68dd080517407fb3a4f05d91fed27d7"
## Size (B):       138428
## Owner:          6a1a843f5cdc0cd887a7117c91d8bb3e0c9d31ba5db7441f7d01d6ade80cfac3
## Storage class:  STANDARD
##
## $Contents
## Key:            CaseStudy2CompSet No Attrition.csv
## LastModified:   2022-07-18T19:01:55.000Z
## ETag:           "6c9d92b8a6fc5fd805ff0a5d4dfddde0"
## Size (B):       47686
## Owner:          6a1a843f5cdc0cd887a7117c91d8bb3e0c9d31ba5db7441f7d01d6ade80cfac3
## Storage class:  STANDARD
##
## $Contents
## Key:            CaseStudy2CompSet No Salary.xlsx
## LastModified:   2022-07-18T19:01:56.000Z
## ETag:           "bdcb211847739638a631f828a3278339"
## Size (B):       56381
## Owner:          6a1a843f5cdc0cd887a7117c91d8bb3e0c9d31ba5db7441f7d01d6ade80cfac3
## Storage class:  STANDARD
```

```
# read and write from ojbect

#Read in Creativity.csv
#1st file
case2predictions = s3read_using(FUN = read.csv,
                    bucket = "smuddsproject2",
                    object = "Case2PredictionsClassifyEXAMPLE.csv")



# 2nd file
RegressEXAMPLE = s3read_using(FUN = read.csv,
                    bucket = "smuddsproject2",
                    object = "Case2PredictionsRegressEXAMPLE.csv")




# 3rd file

Casestudy2 = s3read_using(FUN = read.csv,
                    bucket = "smuddsproject2",
                    object = "CaseStudy2-data.csv")



# 4th  file

Casestudy2NoA = s3read_using(FUN = read.csv,
                    bucket = "smuddsproject2",
                    object = "CaseStudy2CompSet No Attrition.csv")




# 5th  file

Casestudy2NoS = s3read_using(FUN = read_xlsx,
                    bucket = "smuddsproject2",
                    object = "CaseStudy2CompSet No Salary.xlsx")
```

Data Summary

```
data <- Casestudy2
summary(data)
```

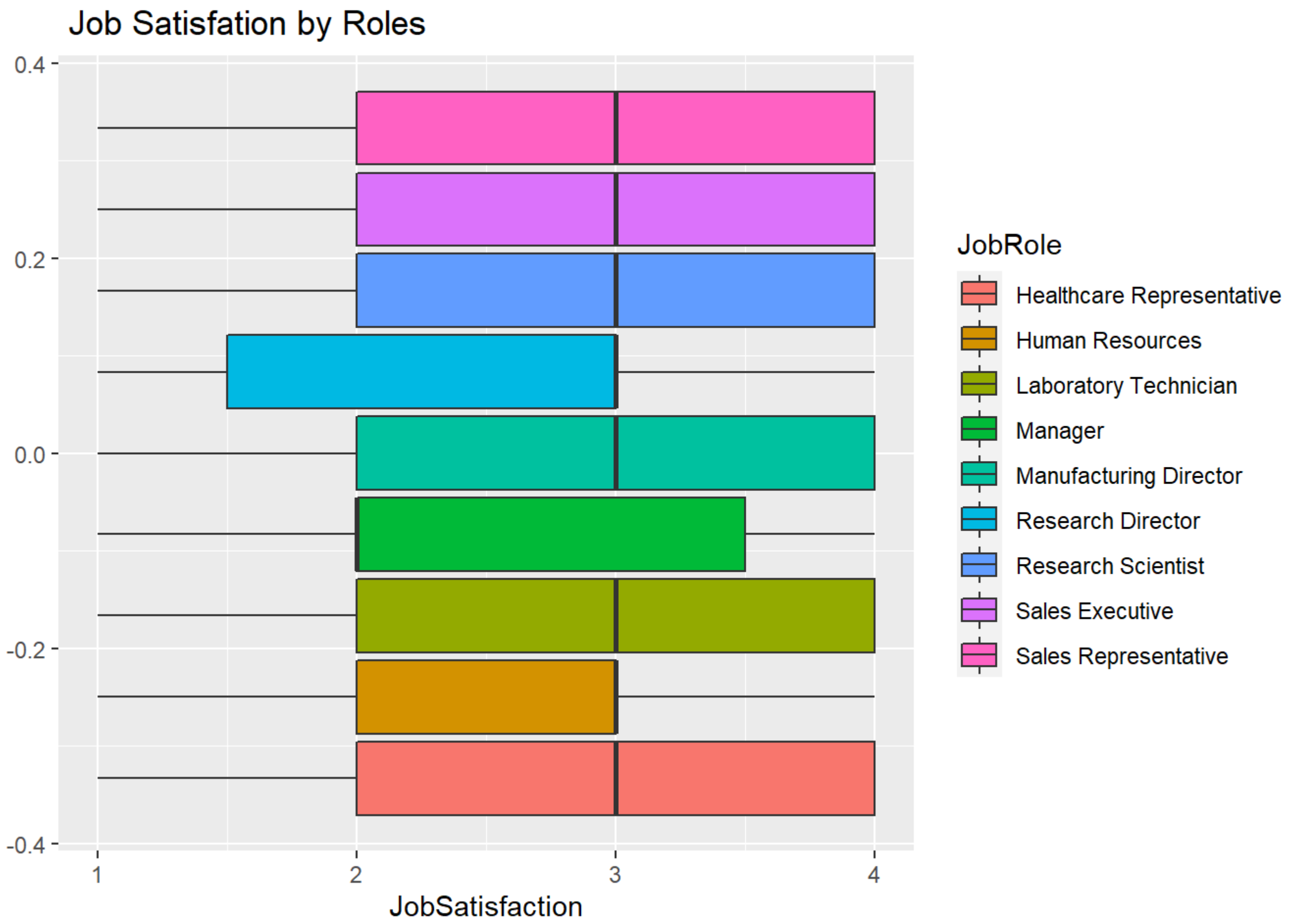
```
##        ID             Age         Attrition         BusinessTravel    
##  Min.   :  1.0   Min.   :18.00   Length:870         Length:870        
##  1st Qu.:218.2   1st Qu.:30.00   Class :character   Class :character  
##  Median :435.5   Median :35.00   Mode  :character   Mode  :character  
##  Mean   :435.5   Mean   :36.83                                        
##  3rd Qu.:652.8   3rd Qu.:43.00                                        
##  Max.   :870.0   Max.   :60.00                                        
##    DailyRate       Department        DistanceFromHome   Education    
##  Min.   : 103.0   Length:870         Min.   : 1.000   Min.   :1.000  
##  1st Qu.: 472.5   Class :character   1st Qu.: 2.000   1st Qu.:2.000  
##  Median : 817.5   Mode  :character   Median : 7.000   Median :3.000  
##  Mean   : 815.2                      Mean   : 9.339   Mean   :2.901  
##  3rd Qu.:1165.8                      3rd Qu.:14.000   3rd Qu.:4.000  
##  Max.   :1499.0                      Max.   :29.000   Max.   :5.000  
##  EducationField     EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Length:870         Min.   :1     Min.   :   1.0   Min.   :1.000          
##  Class :character   1st Qu.:1     1st Qu.: 477.2   1st Qu.:2.000          
##  Mode  :character   Median :1     Median :1039.0   Median :3.000          
##                     Mean   :1     Mean   :1029.8   Mean   :2.701          
##                     3rd Qu.:1     3rd Qu.:1561.5   3rd Qu.:4.000          
##                     Max.   :1     Max.   :2064.0   Max.   :4.000          
##     Gender            HourlyRate     JobInvolvement     JobLevel    
##  Length:870         Min.   : 30.00   Min.   :1.000   Min.   :1.000  
##  Class :character   1st Qu.: 48.00   1st Qu.:2.000   1st Qu.:1.000  
##  Mode  :character   Median : 66.00   Median :3.000   Median :2.000  
##                     Mean   : 65.61   Mean   :2.723   Mean   :2.039  
##                     3rd Qu.: 83.00   3rd Qu.:3.000   3rd Qu.:3.000  
##                     Max.   :100.00   Max.   :4.000   Max.   :5.000  
##    JobRole          JobSatisfaction MaritalStatus      MonthlyIncome  
##  Length:870         Min.   :1.000   Length:870         Min.   : 1081  
##  Class :character   1st Qu.:2.000   Class :character   1st Qu.: 2840  
##  Mode  :character   Median :3.000   Mode  :character   Median : 4946  
##                     Mean   :2.709                      Mean   : 6390  
##                     3rd Qu.:4.000                      3rd Qu.: 8182  
##                     Max.   :4.000                      Max.   :19999  
##   MonthlyRate    NumCompaniesWorked    Over18            OverTime        
##  Min.   : 2094   Min.   :0.000      Length:870         Length:870        
##  1st Qu.: 8092   1st Qu.:1.000      Class :character   Class :character  
##  Median :14074   Median :2.000      Mode  :character   Mode  :character  
##  Mean   :14326   Mean   :2.728  
##  3rd Qu.:20456   3rd Qu.:4.000  
##  Max.   :26997   Max.   :9.000  
##  PercentSalaryHike PerformanceRating RelationshipSatisfaction StandardHours
##  Min.   :11.0      Min.   :3.000     Min.   :1.000            Min.   :80   
##  1st Qu.:12.0      1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80   
##  Median :14.0      Median :3.000     Median :3.000            Median :80   
##  Mean   :15.2      Mean   :3.152     Mean   :2.707            Mean   :80   
##  3rd Qu.:18.0      3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80   
##  Max.   :25.0      Max.   :4.000     Max.   :4.000            Max.   :80   
##  StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
##  Min.   :0.0000   Min.   : 0.00     Min.   :0.000         Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000  
##  Median :1.0000   Median :10.00     Median :3.000         Median :3.000  
##  Mean   :0.7839   Mean   :11.05     Mean   :2.832         Mean   :2.782  
##  3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000  
##  Max.   :3.0000   Max.   :40.00     Max.   :6.000         Max.   :4.000  
##  YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion
##  Min.   : 0.000   Min.   : 0.000     Min.   : 0.000         
##  1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000         
##  Median : 5.000   Median : 3.000     Median : 1.000         
##  Mean   : 6.962   Mean   : 4.205     Mean   : 2.169         
##  3rd Qu.:10.000   3rd Qu.: 7.000     3rd Qu.: 3.000         
##  Max.   :40.000   Max.   :18.000     Max.   :15.000         
##  YearsWithCurrManager
##  Min.   : 0.00       
##  1st Qu.: 2.00       
##  Median : 3.00       
##  Mean   : 4.14       
##  3rd Qu.: 7.00       
##  Max.   :17.00
```

```
# change multiple columns to factors
data[c(3,4,6,9,13,17,19,23,24)] <- lapply(data[c(3,4,6,9,13,17,19,23,24)],as.factor)

summary(data)
```

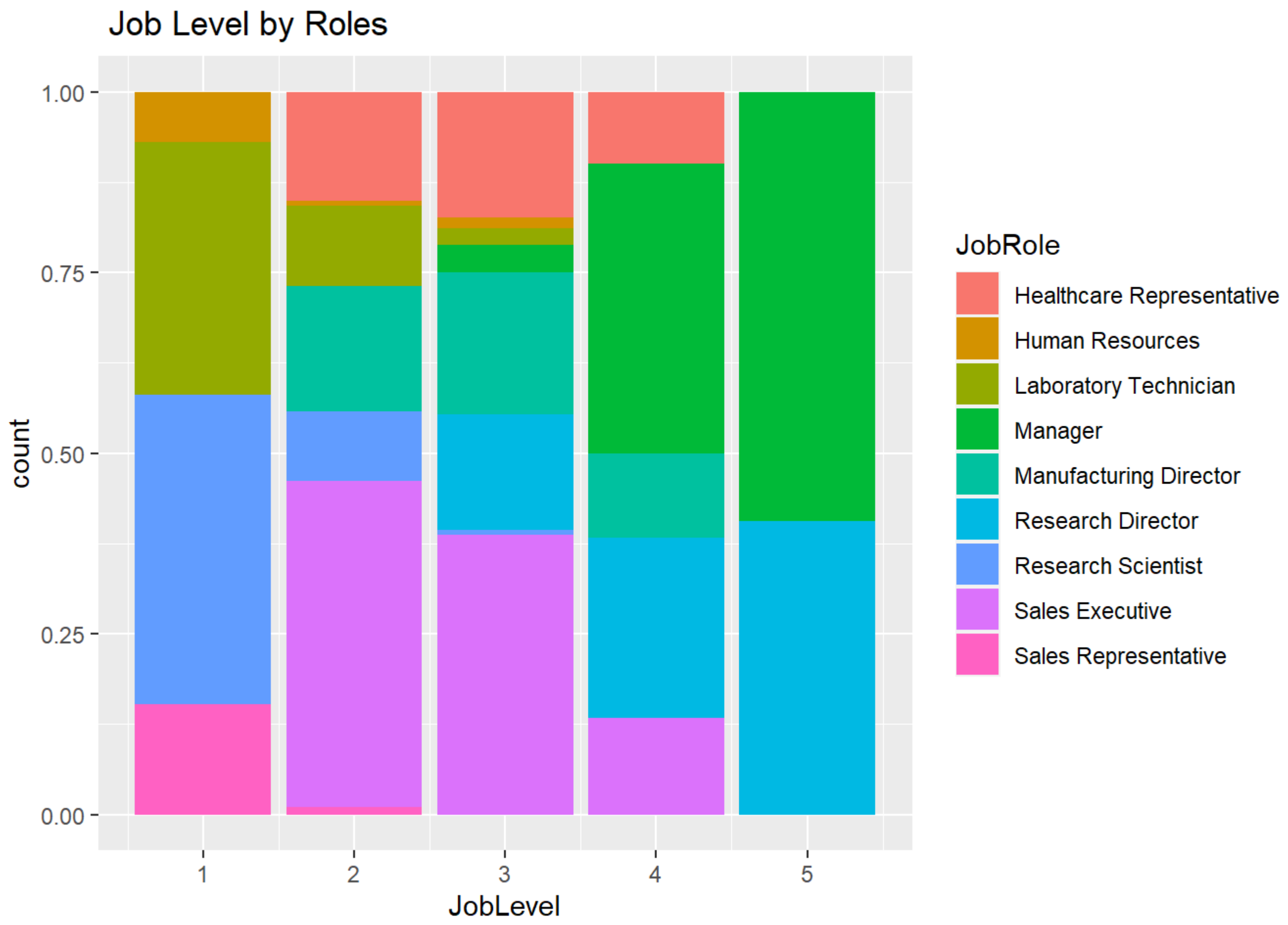
```
##        ID             Age        Attrition           BusinessTravel
##  Min.   :  1.0   Min.   :18.00   No :730   Non-Travel       : 94
##  1st Qu.:218.2   1st Qu.:30.00   Yes:140   Travel_Frequently:158
##  Median :435.5   Median :35.00             Travel_Rarely    :618
##  Mean   :435.5   Mean   :36.83
##  3rd Qu.:652.8   3rd Qu.:43.00
##  Max.   :870.0   Max.   :60.00
##
##    DailyRate                       Department  DistanceFromHome   Education
##  Min.   : 103.0   Human Resources       : 35   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 472.5   Research & Development:562   1st Qu.: 2.000   1st Qu.:2.000
##  Median : 817.5   Sales                 :273   Median : 7.000   Median :3.000
##  Mean   : 815.2                                Mean   : 9.339   Mean   :2.901
##  3rd Qu.:1165.8                                3rd Qu.:14.000   3rd Qu.:4.000
##  Max.   :1499.0                                Max.   :29.000   Max.   :5.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 15    Min.   :1     Min.   :   1.0   Min.   :1.000
##  Life Sciences   :358    1st Qu.:1     1st Qu.: 477.2   1st Qu.:2.000
##  Marketing       :100    Median :1     Median :1039.0   Median :3.000
##  Medical         :270    Mean   :1     Mean   :1029.8   Mean   :2.701
##  Other           : 52    3rd Qu.:1     3rd Qu.:1561.5   3rd Qu.:4.000
##  Technical Degree: 75    Max.   :1     Max.   :2064.0   Max.   :4.000
##
##     Gender      HourlyRate     JobInvolvement     JobLevel
##  Female:354   Min.   : 30.00   Min.   :1.000   Min.   :1.000
##  Male  :516   1st Qu.: 48.00   1st Qu.:2.000   1st Qu.:1.000
##               Median : 66.00   Median :3.000   Median :2.000
##               Mean   : 65.61   Mean   :2.723   Mean   :2.039
##               3rd Qu.: 83.00   3rd Qu.:3.000   3rd Qu.:3.000
##               Max.   :100.00   Max.   :4.000   Max.   :5.000
##
##                       JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
##  Sales Executive          :200   Min.   :1.000   Divorced:191   Min.   : 1081
##  Research Scientist       :172   1st Qu.:2.000   Married :410   1st Qu.: 2840
##  Laboratory Technician    :153   Median :3.000   Single  :269   Median : 4946
##  Manufacturing Director   : 87   Mean   :2.709                  Mean   : 6390
##  Healthcare Representative: 76   3rd Qu.:4.000                  3rd Qu.: 8182
##  Sales Representative     : 53   Max.   :4.000                  Max.   :19999
##  (Other)                  :129                                              
##   MonthlyRate    NumCompaniesWorked Over18  OverTime  PercentSalaryHike
##  Min.   : 2094   Min.   :0.000      Y:870   No :618   Min.   :11.0     
##  1st Qu.: 8092   1st Qu.:1.000              Yes:252   1st Qu.:12.0     
##  Median :14074   Median :2.000                        Median :14.0     
##  Mean   :14326   Mean   :2.728                        Mean   :15.2     
##  3rd Qu.:20456   3rd Qu.:4.000                        3rd Qu.:18.0     
##  Max.   :26997   Max.   :9.000                        Max.   :25.0     
##                                                                        
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000  
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000  
##  Median :3.000     Median :3.000            Median :80    Median :1.0000  
##  Mean   :3.152     Mean   :2.707            Mean   :80    Mean   :0.7839  
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000  
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000  
##                                                                           
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany  
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000  
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000  
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000  
##  Mean   :11.05     Mean   :2.832         Mean   :2.782   Mean   : 6.962  
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.:10.000  
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000  
##                                                                          
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.00       
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.00       
##  Median : 3.000     Median : 1.000          Median : 3.00       
##  Mean   : 4.205     Mean   : 2.169          Mean   : 4.14       
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.00       
##  Max.   :18.000     Max.   :15.000          Max.   :17.00       
## 
```

Data Analysis

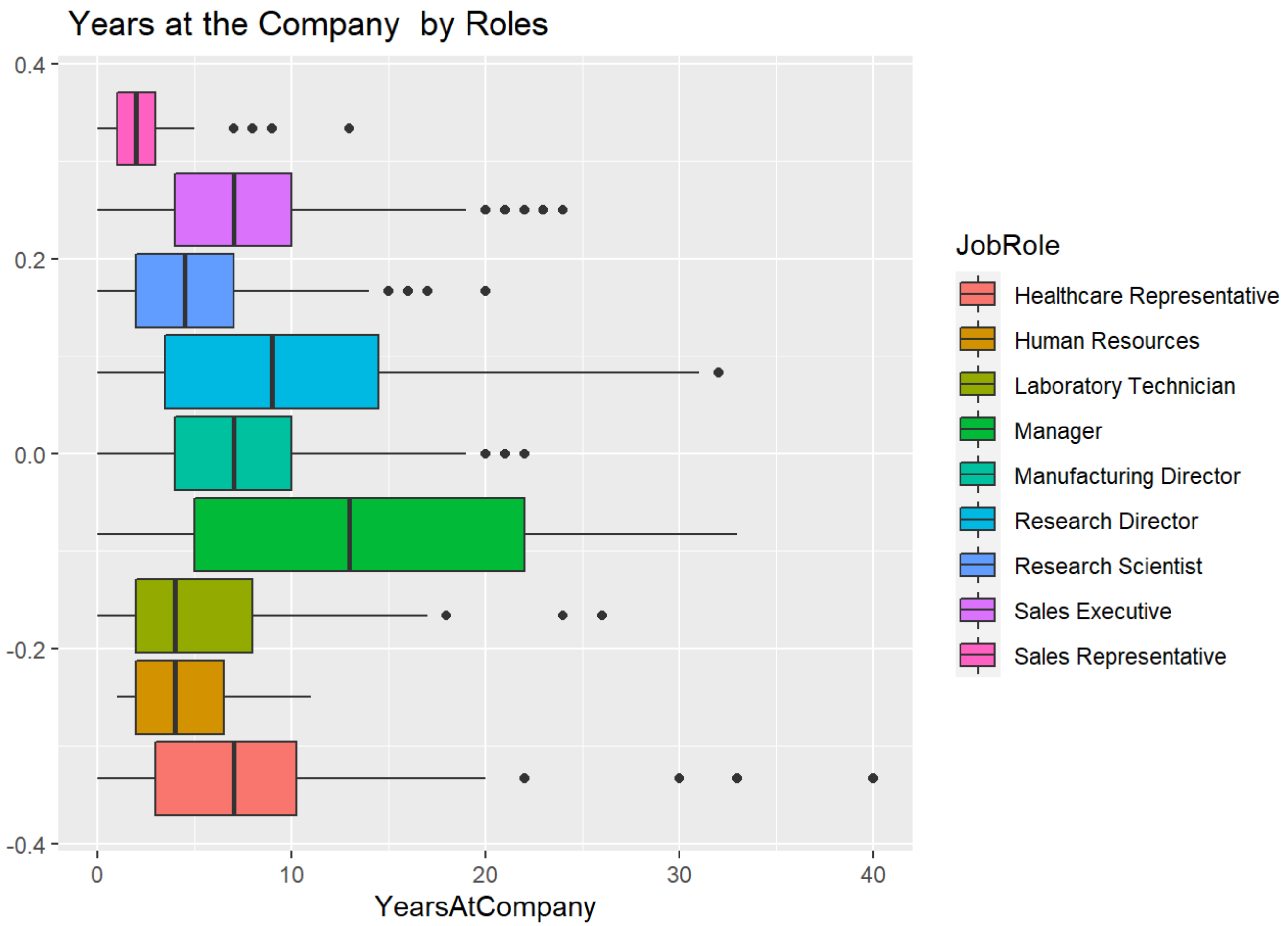
```
# 1) The median job satisfaction is the same for all roles. However, Human Resources and Research Director roles don't get t
o the highest level of 4.
data %>% ggplot(aes(x= JobSatisfaction, fill=JobRole)) + geom_boxplot() + ggtitle(" Job Satisfation by Roles")
```



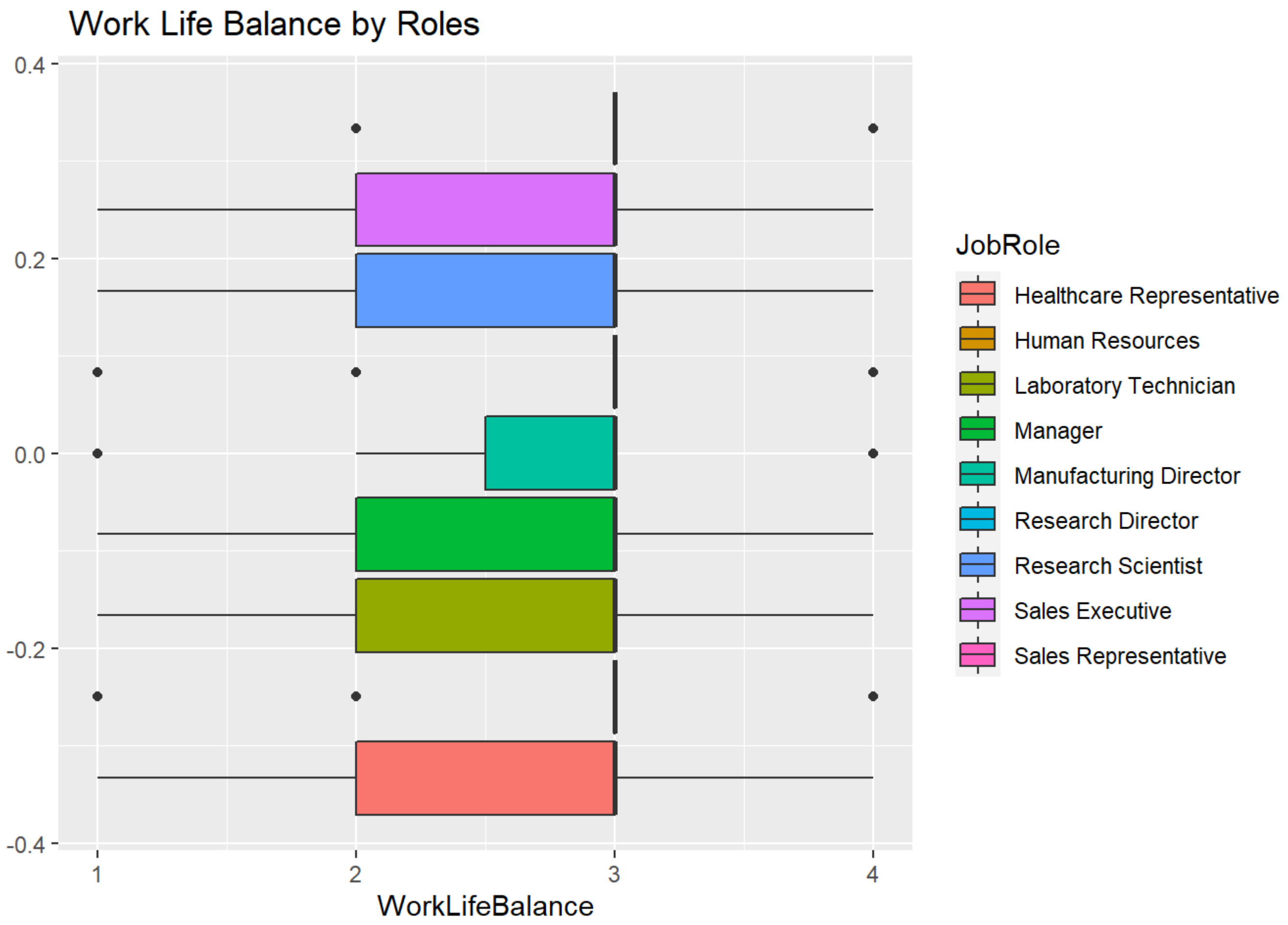
```
# 2) Most of the higher level jobs(4,5) are occupied by Managers and Research Directors with few from Healthcare and Sales E
xecutive.
data %>% ggplot(aes(x= JobLevel, fill=JobRole)) + geom_bar(position = "fill") + ggtitle(" Job Level by Roles")
```



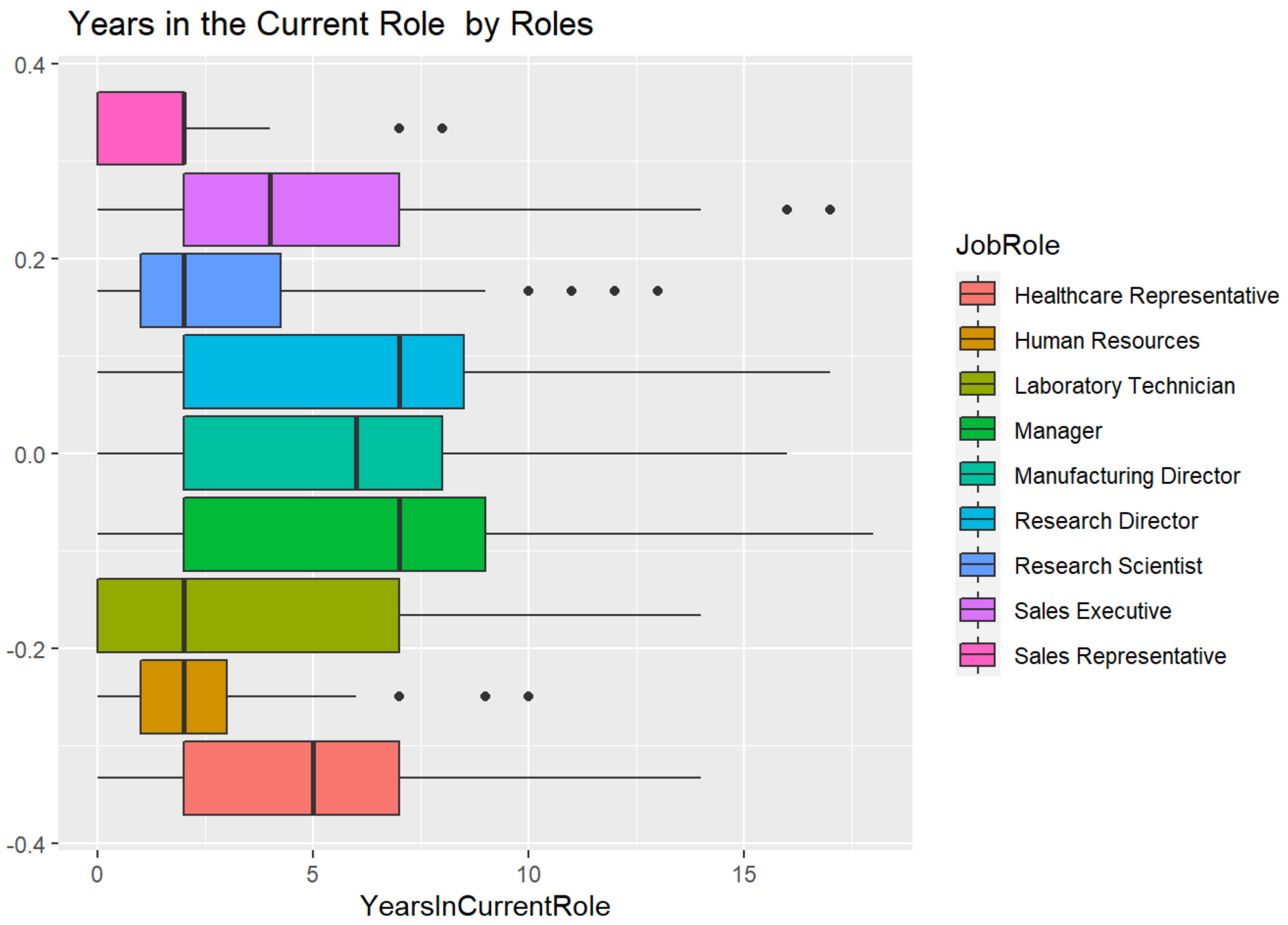
```
# 3) Managers stay at the company the longest followed by the Research Directors.This can be attributed to the higher job ro
les discussed previously in 2 for these positions.
data %>% ggplot(aes(x= YearsAtCompany, fill=JobRole)) + geom_boxplot() + ggtitle(" Years at the Company  by Roles")
```



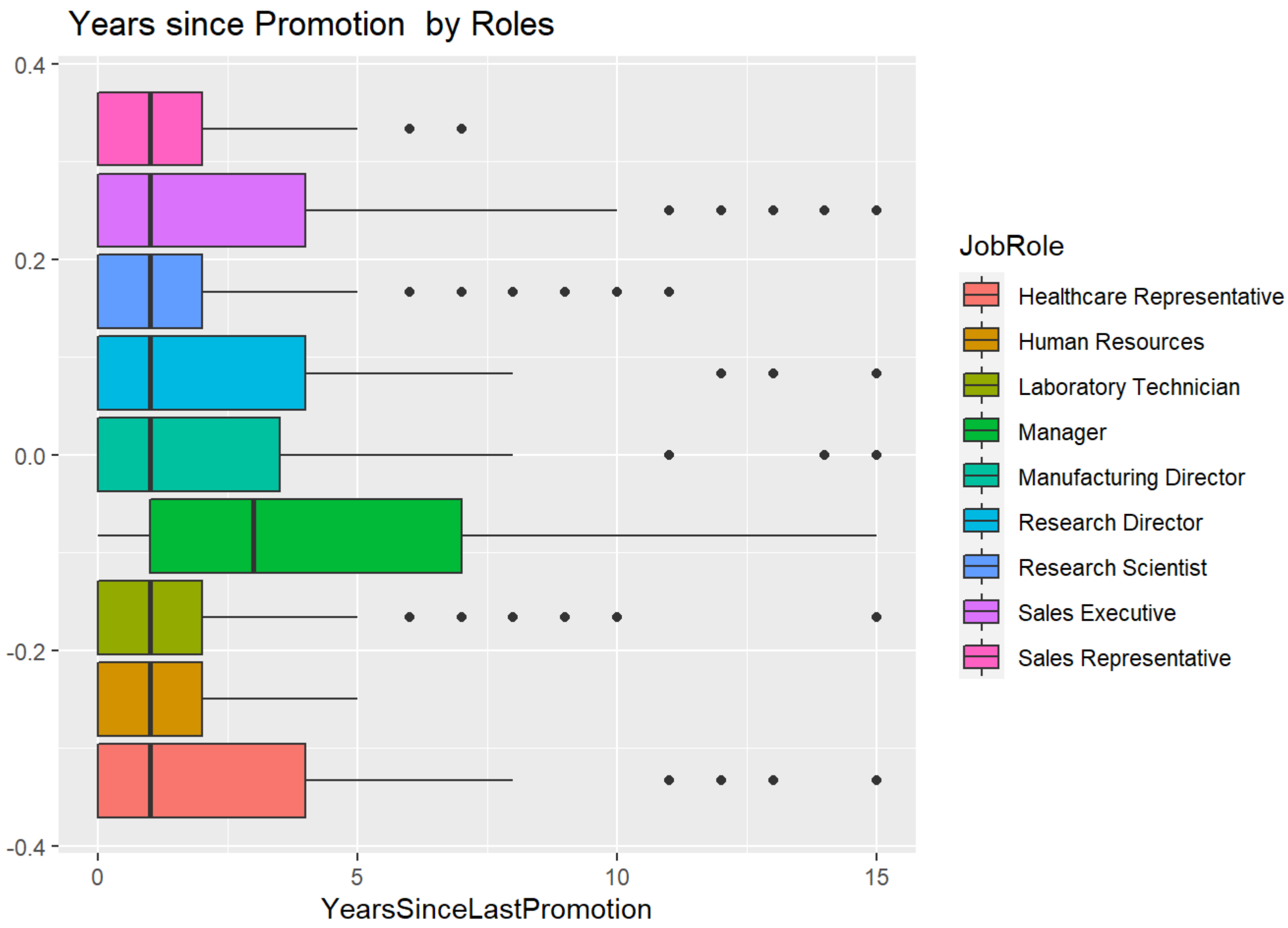
```
# 4) All roles have same median worklife balance.
data %>% ggplot(aes(x= WorkLifeBalance, fill=JobRole)) + geom_boxplot() + ggtitle(" Work Life Balance by Roles")
```



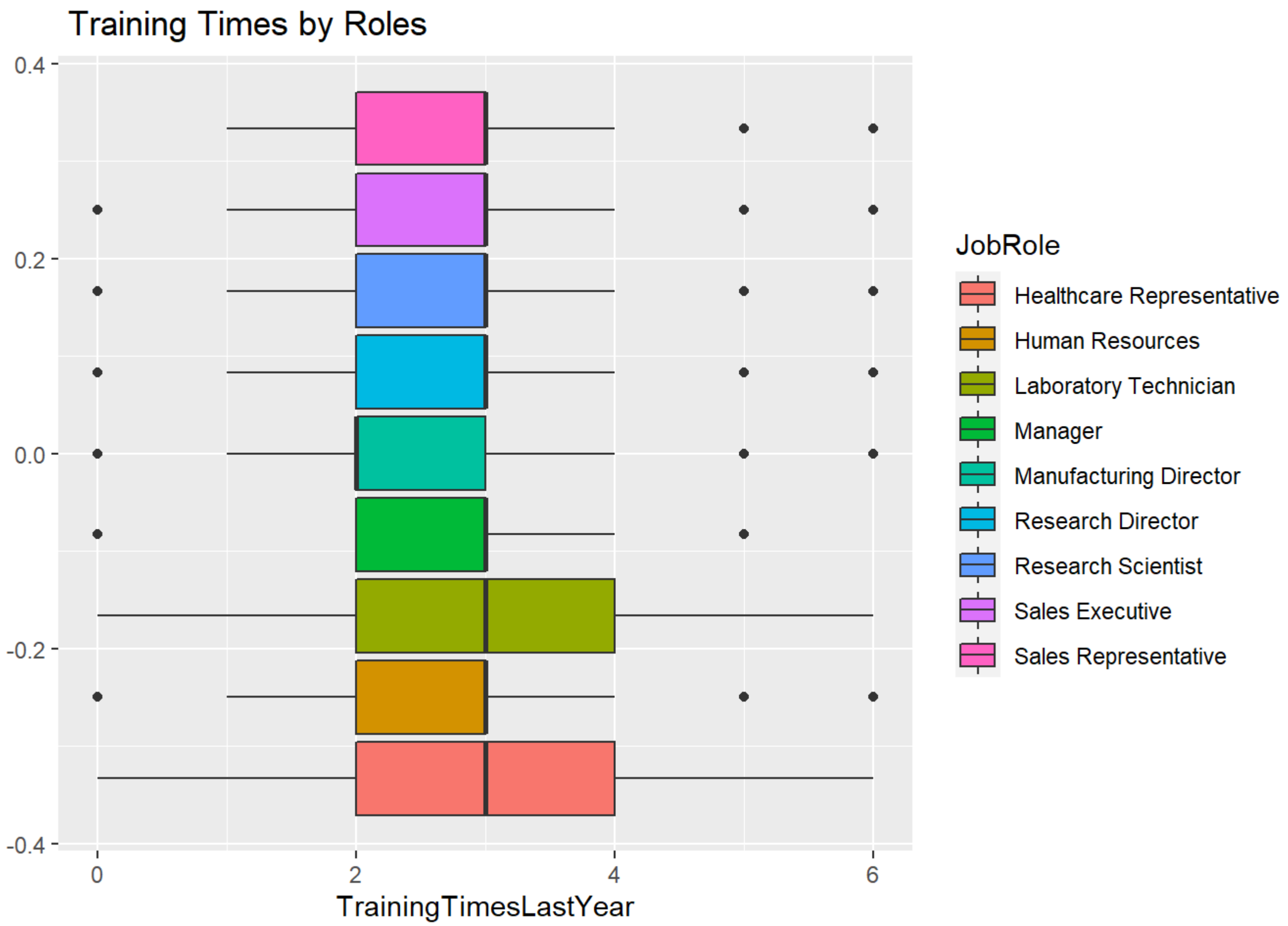
```
# 5) Managers and Research Directors stay longer in their roles whereas Human Resources and Sales Representative stay the le
ast.
data %>% ggplot(aes(x= YearsInCurrentRole , fill=JobRole)) + geom_boxplot()+ ggtitle(" Years in the Current Role  by Roles")
```



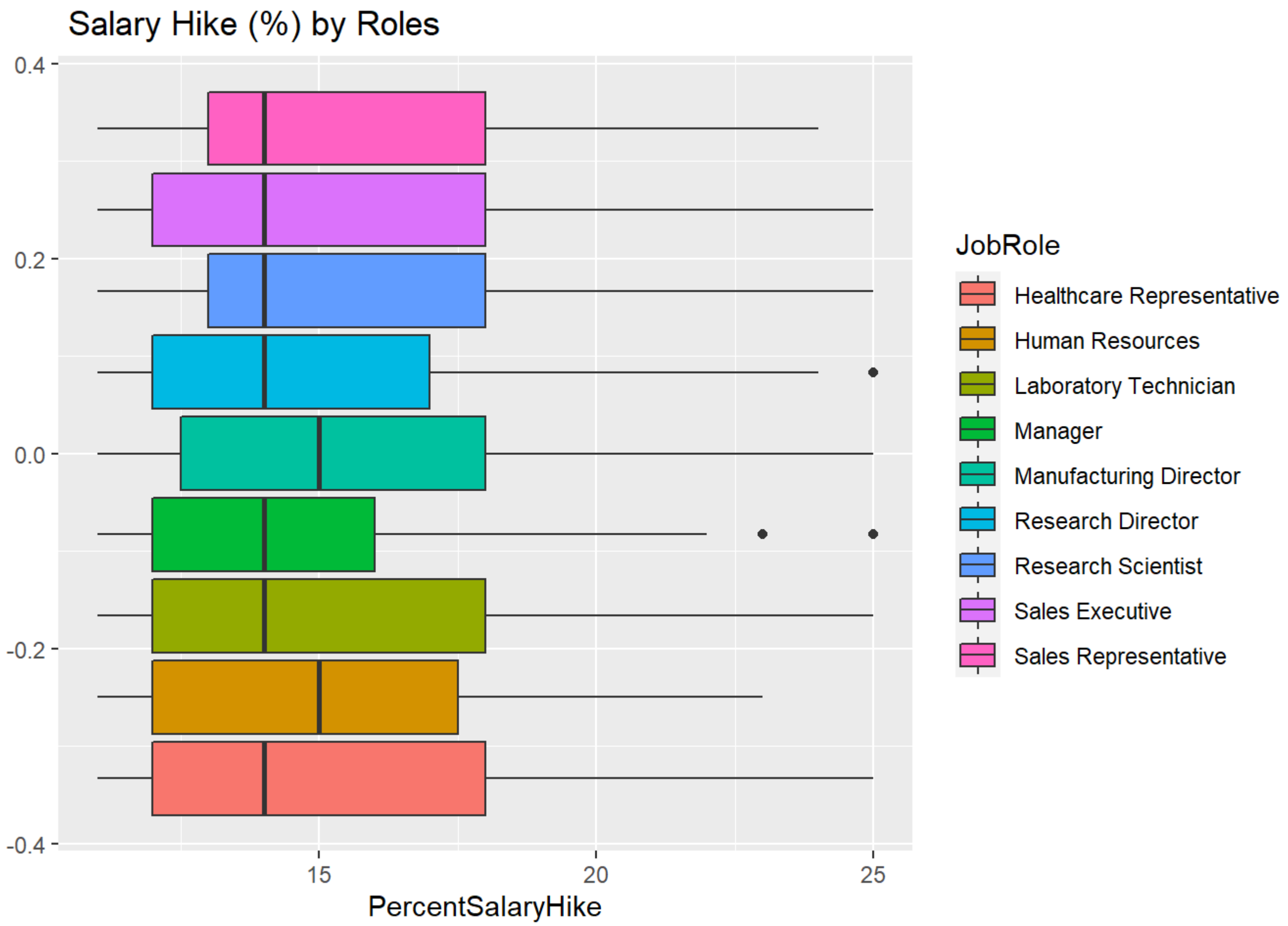
```
# 6) Mangers have the most median years since promotion. Most managers are in higher level positions which is is possibly th
e reason that they get promoted less frequently.
data %>% ggplot(aes(x= YearsSinceLastPromotion , fill=JobRole)) + geom_boxplot()+ ggtitle(" Years since Promotion  by Role
s")
```



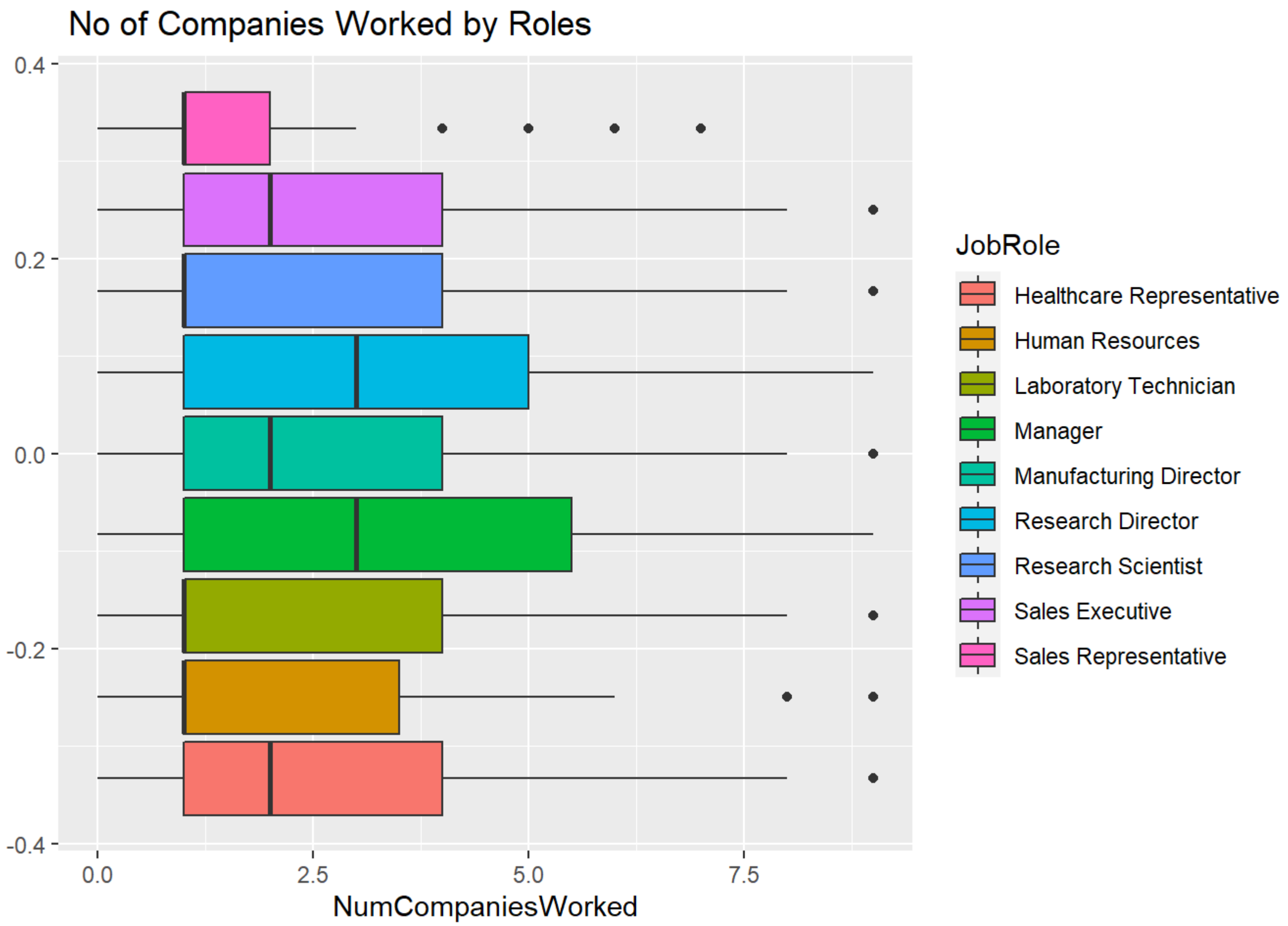
```
# 7) Training times were similar for most positions. Manufacturing Directors median training is ~17% less that other roles.
data %>% ggplot(aes(x= TrainingTimesLastYear , fill=JobRole)) + geom_boxplot() + ggtitle(" Training Times by Roles")
```



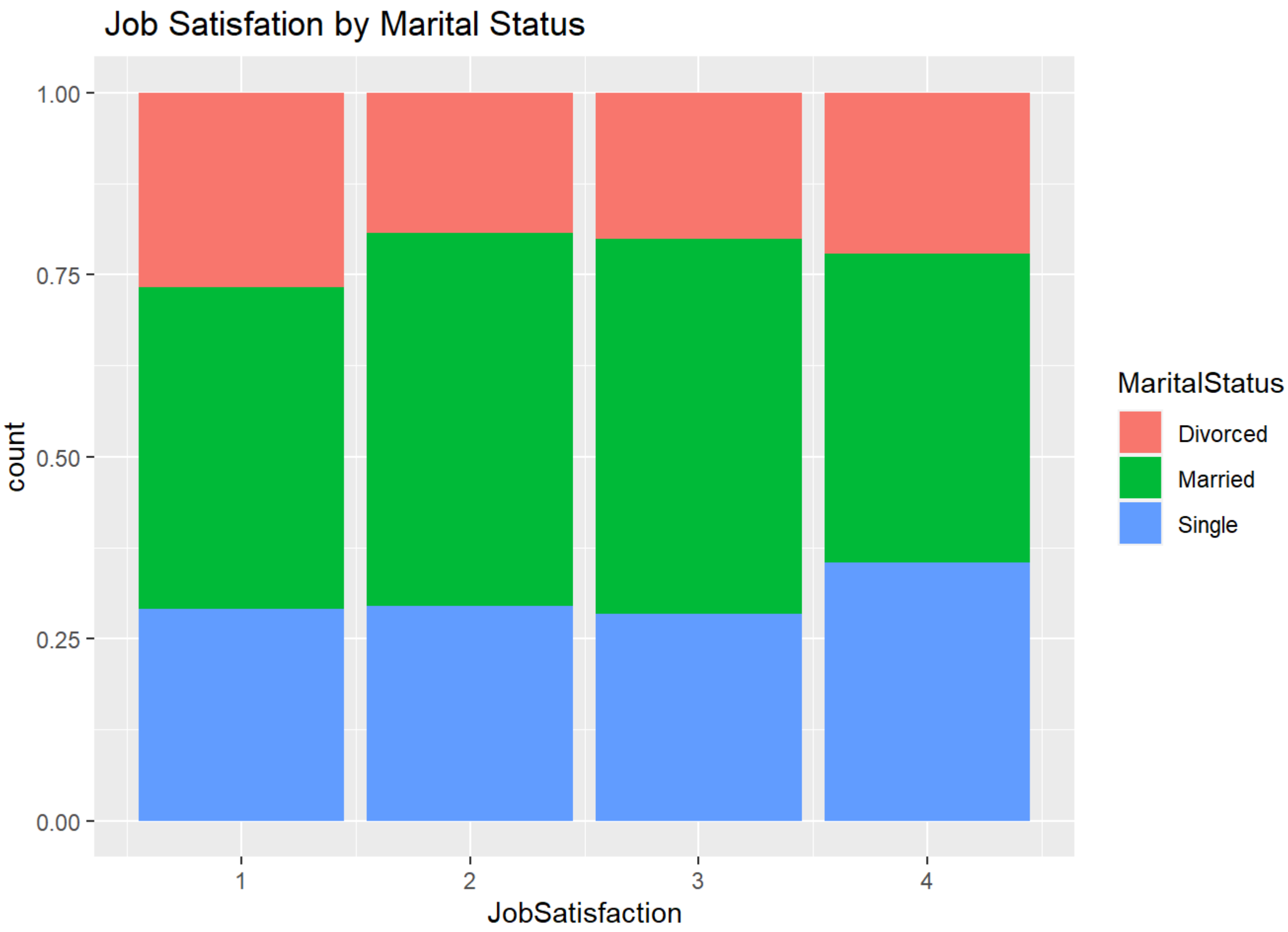
```
# 8) Manufacturing Director and Human Resources had the most median salary hikes in percent.
data %>% ggplot(aes(x= PercentSalaryHike , fill=JobRole)) + geom_boxplot() + ggtitle(" Salary Hike (%) by Roles")
```



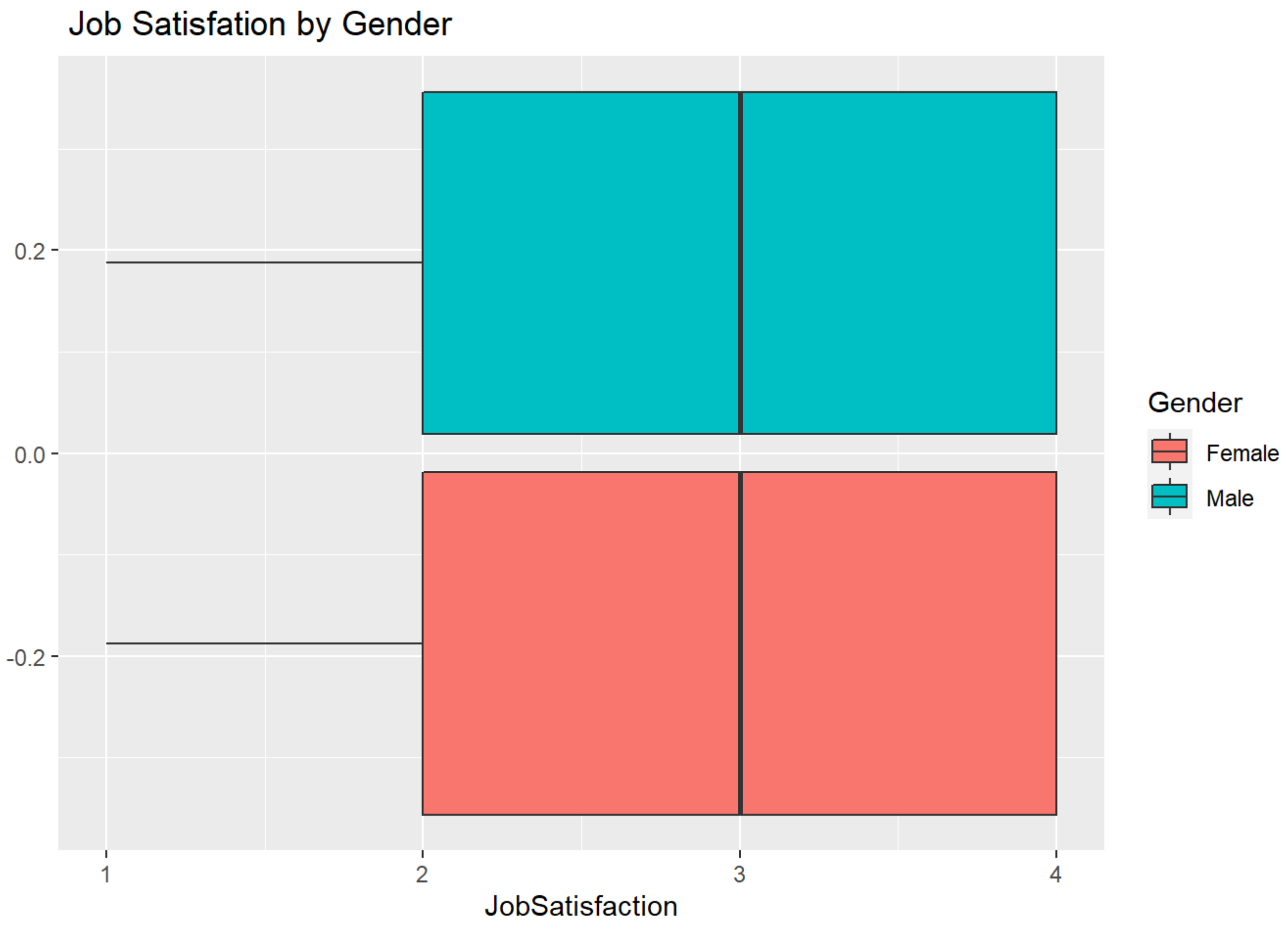
```
# 9) It is interesting that Research Directors and Managers have worked in most companies. They also stay longer in their roles. They most likely bring a lot of experience with them and stay with Frito Lay longer because of the higher level position that they occupy.
data %>% ggplot(aes(x= NumCompaniesWorked  , fill=JobRole)) + geom_boxplot() + ggtitle(" No of Companies Worked by Roles")
```



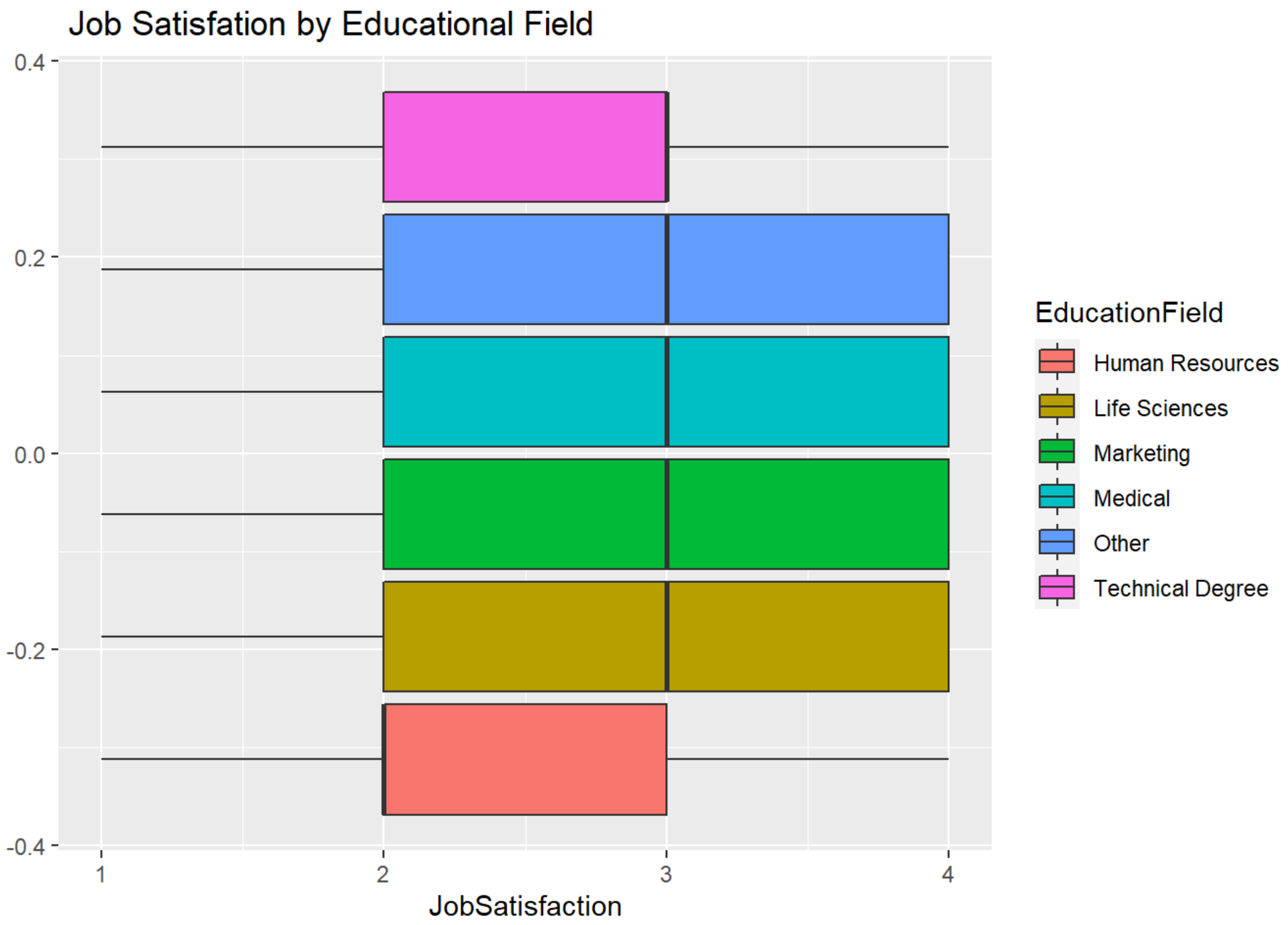
```
# 10)Job Satisfaction is similar among the employes with all marital status.
data %>% ggplot(aes(x= JobSatisfaction, fill=MaritalStatus)) + geom_bar(position="fill") + ggtitle(" Job Satisfation by Mari
tal Status")
```



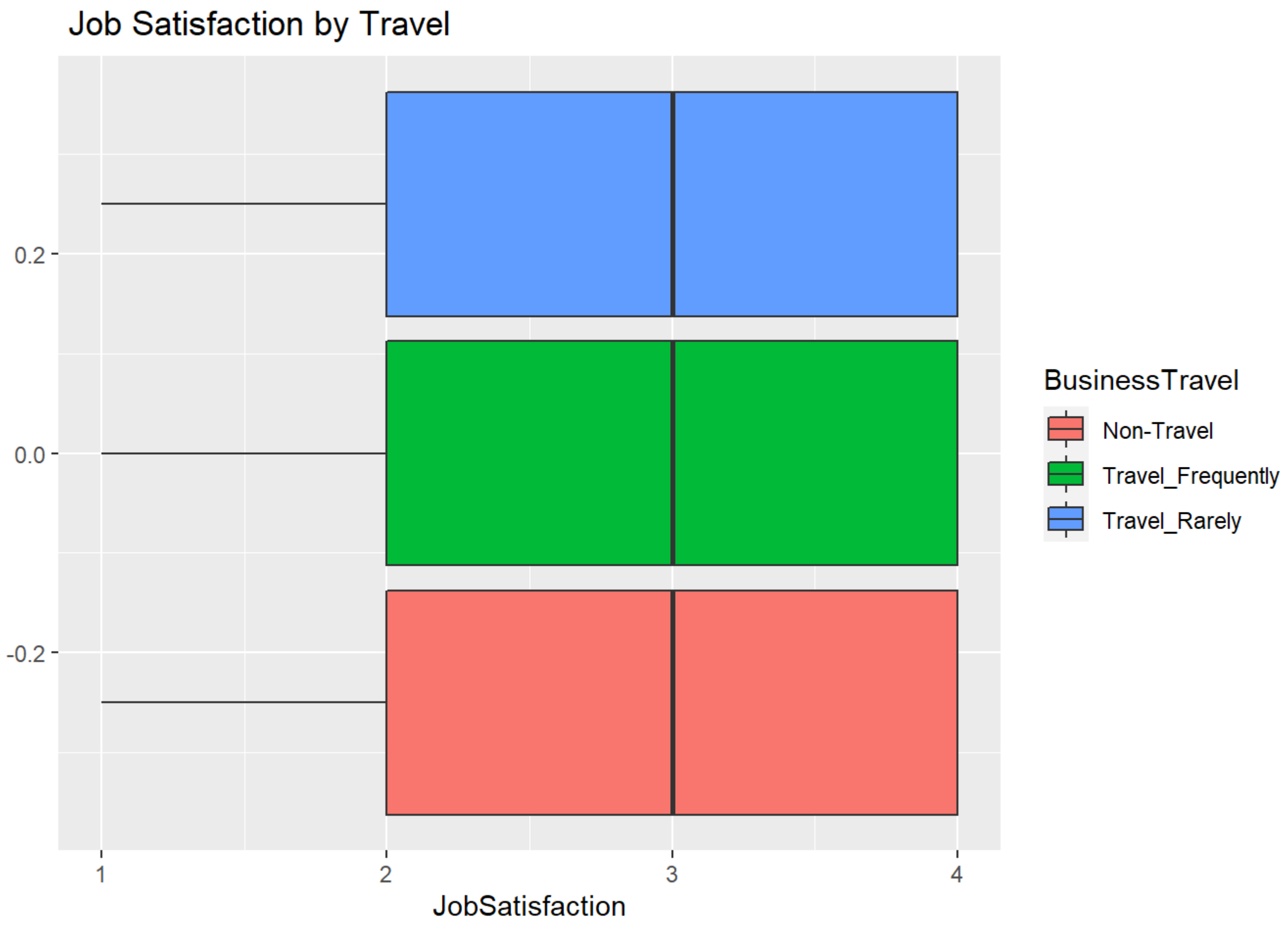
```
# 11) Gender doesn't play a role in job satisfaction.
data %>% ggplot(aes(x= JobSatisfaction, fill=Gender)) + geom_boxplot() + ggtitle(" Job Satisfation by Gender")
```



```
# 12) Human Resources and Technical Degree never get to the highest level of job satisfaction.
data %>% ggplot(aes(x= JobSatisfaction, fill=EducationField)) + geom_boxplot() + ggtitle(" Job Satisfation by Educational Fi
eld")
```

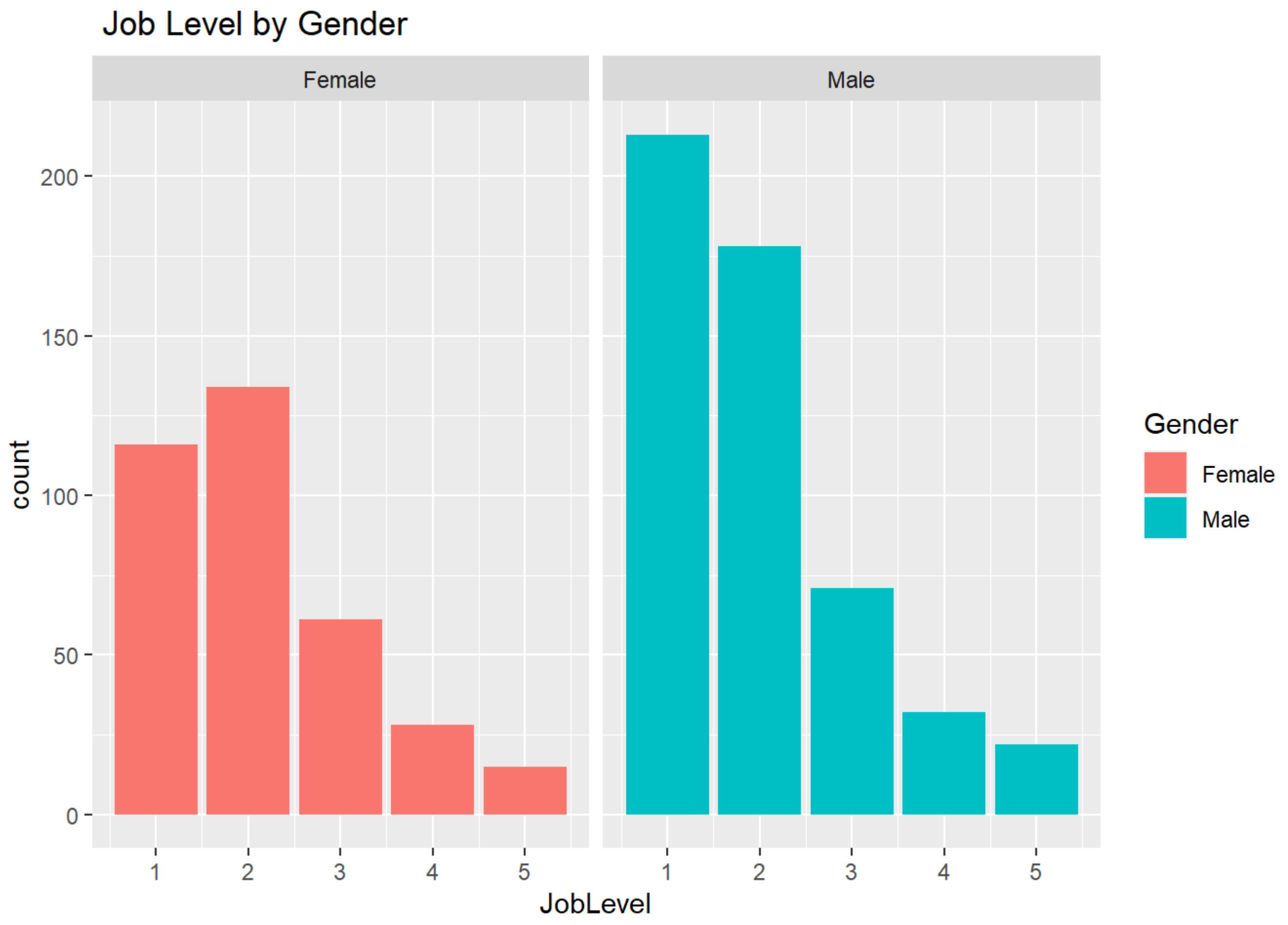


```
# 13) Business travel has no impact on job satisfaction.
data %>% ggplot(aes(x= JobSatisfaction  , fill=BusinessTravel)) + geom_boxplot() + ggtitle(" Job Satisfaction by Travel")
```

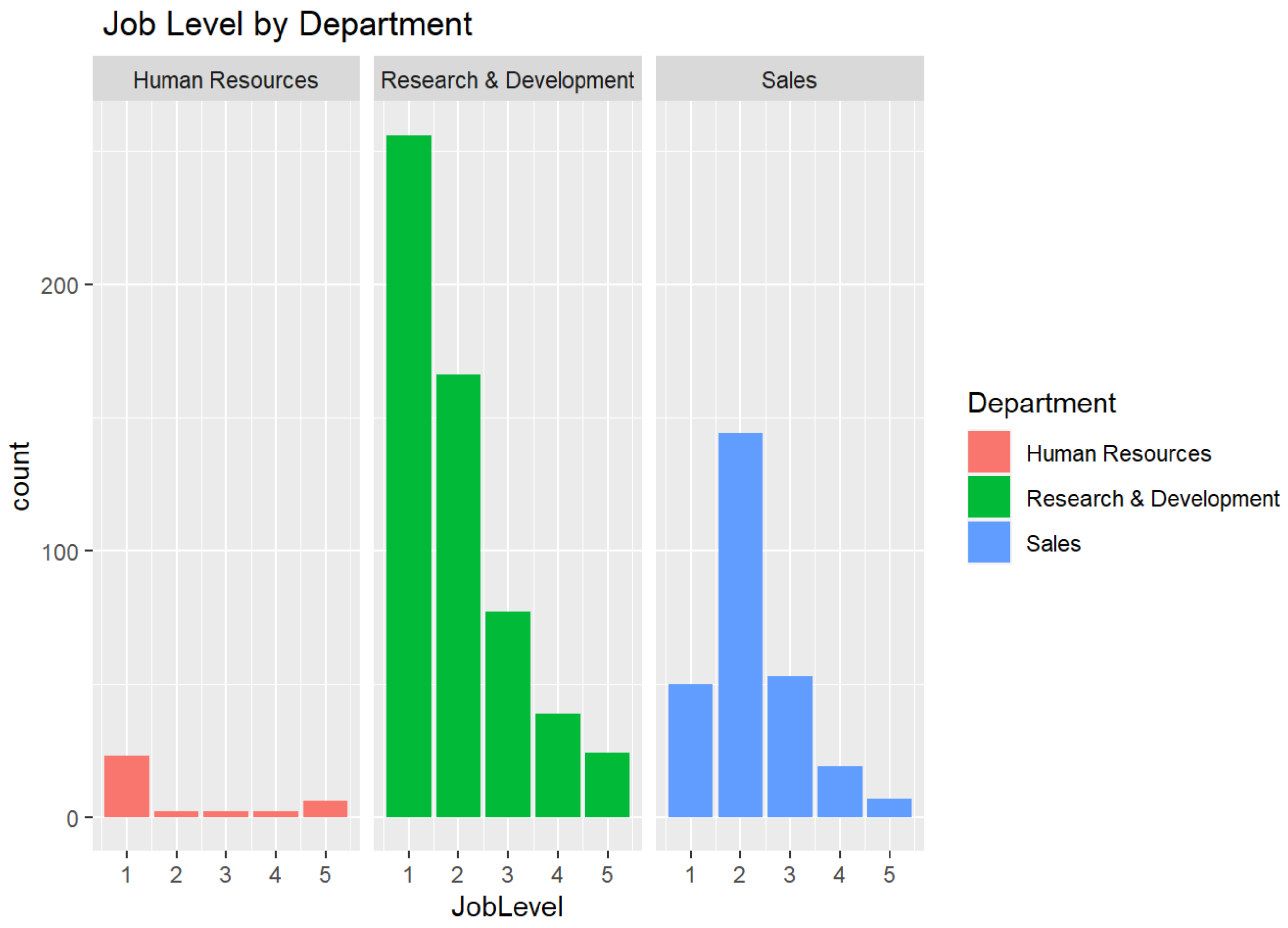


Data Analysis - Job Levels

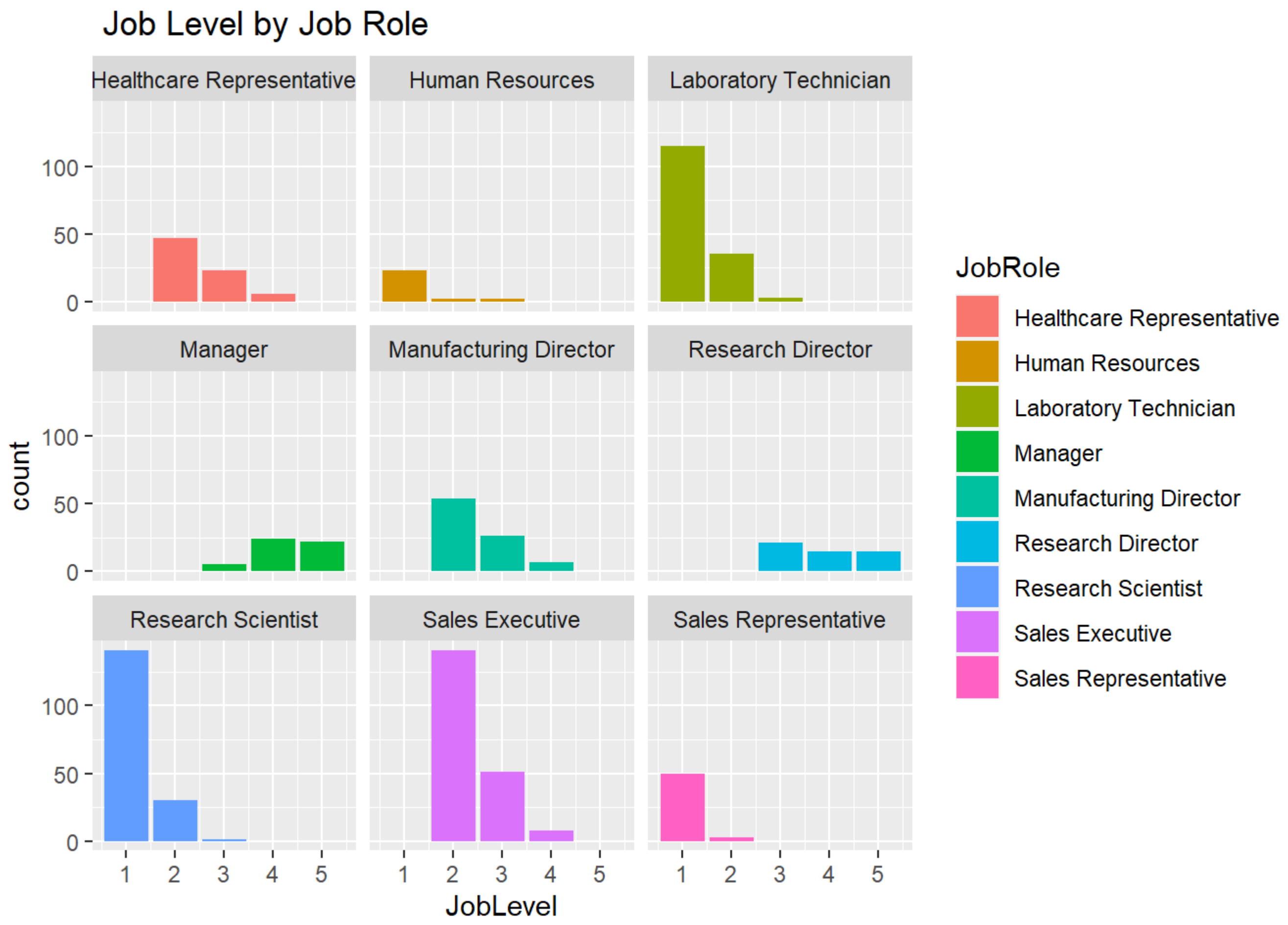
```
#1) There isn't a huge diefference in the higher level jobs between genders. Females are well represented in level 4 and 5 j
obs.
data %>% ggplot(aes(x= JobLevel, fill=Gender)) + geom_bar() + ggtitle(" Job Level by Gender") + facet_wrap(~Gender)
```



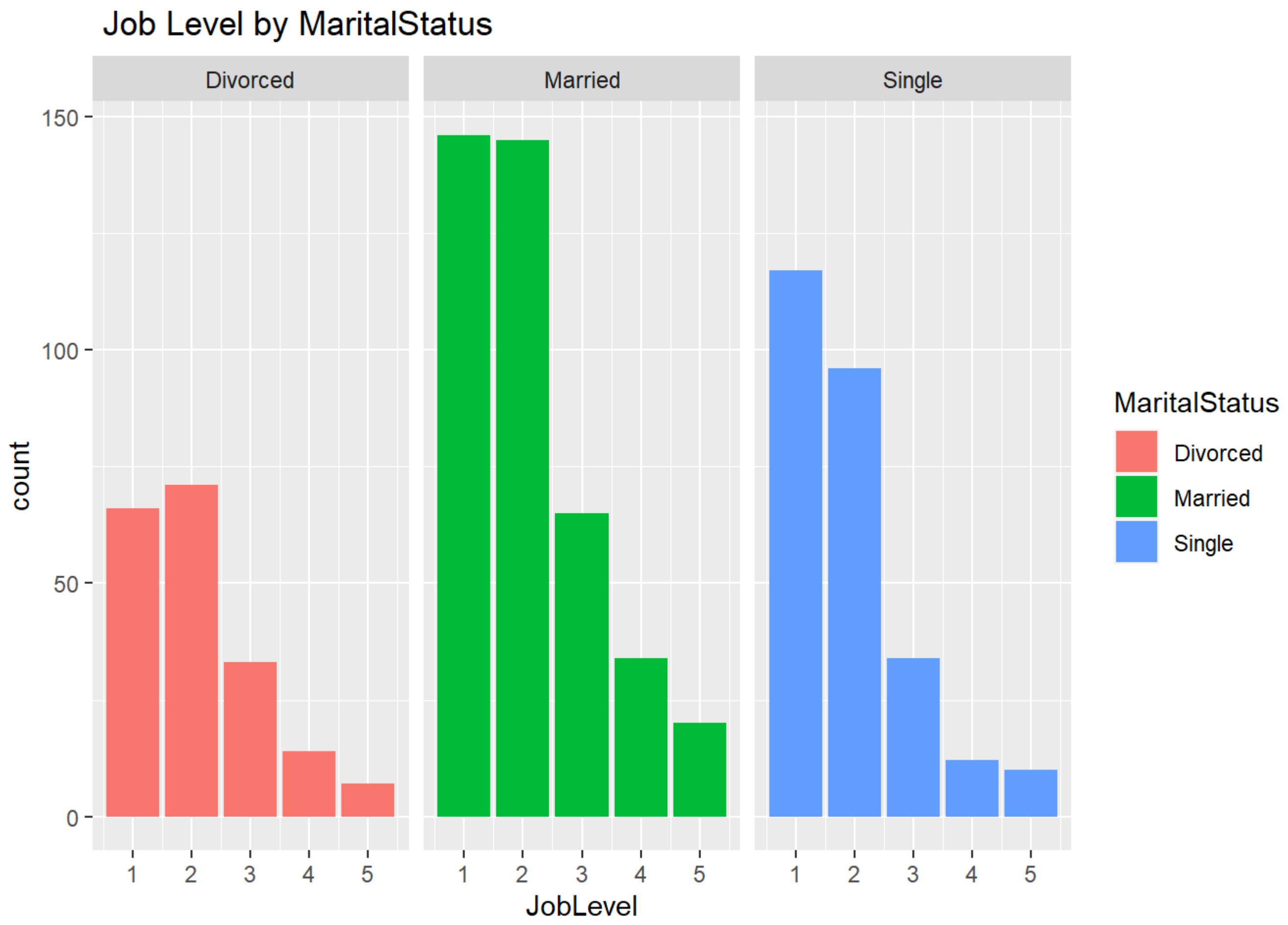
```
#2) Human Resources Department only has lower level roles.
data %>% ggplot(aes(x= JobLevel, fill=Department)) + geom_bar() + ggtitle(" Job Level by Department")+ facet_wrap(~Department)
```



```
#3) Almost all the hih=gher level jobs are filkled by Research Directors and Managers.
data %>% ggplot(aes(x= JobLevel, fill=JobRole)) + geom_bar() + ggtitle(" Job Level by Job Role") + facet_wrap(~JobRole)
```



```
# 4)There isn't a huge difference in job level among the different marital status.
data %>% ggplot(aes(x= JobLevel, fill=MaritalStatus)) + geom_bar() + ggtitle(" Job Level by MaritalStatus") + facet_wrap(~Ma
ritalStatus)
```



Data Analysis - Attrition Visualization

Based on the visualization, we can see that Age, Business Travel, Distance from Home, Job Level, Monthly Income, Stock Option Level,Total Working Years, Years at Company, Years under Current Manager are important varibles that may predict Attrition. We will validate this numerically in the next section.

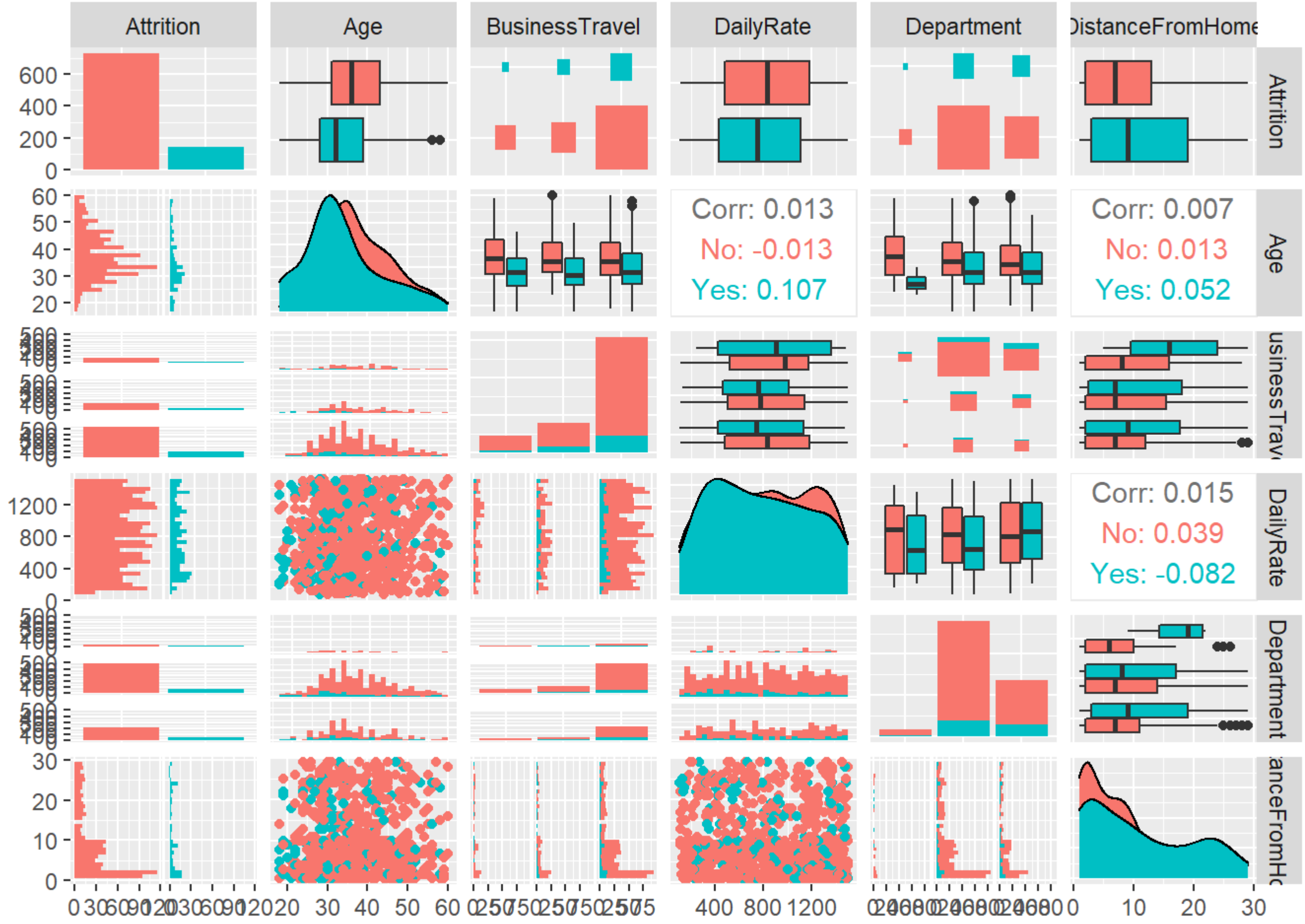
```
library(dplyr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
#Visualizing data with attrition using 5 variables at a time
data %>%  select(Attrition,Age,BusinessTravel,DailyRate,Department,DistanceFromHome) %>% ggpairs(aes(color = Attrition))
```

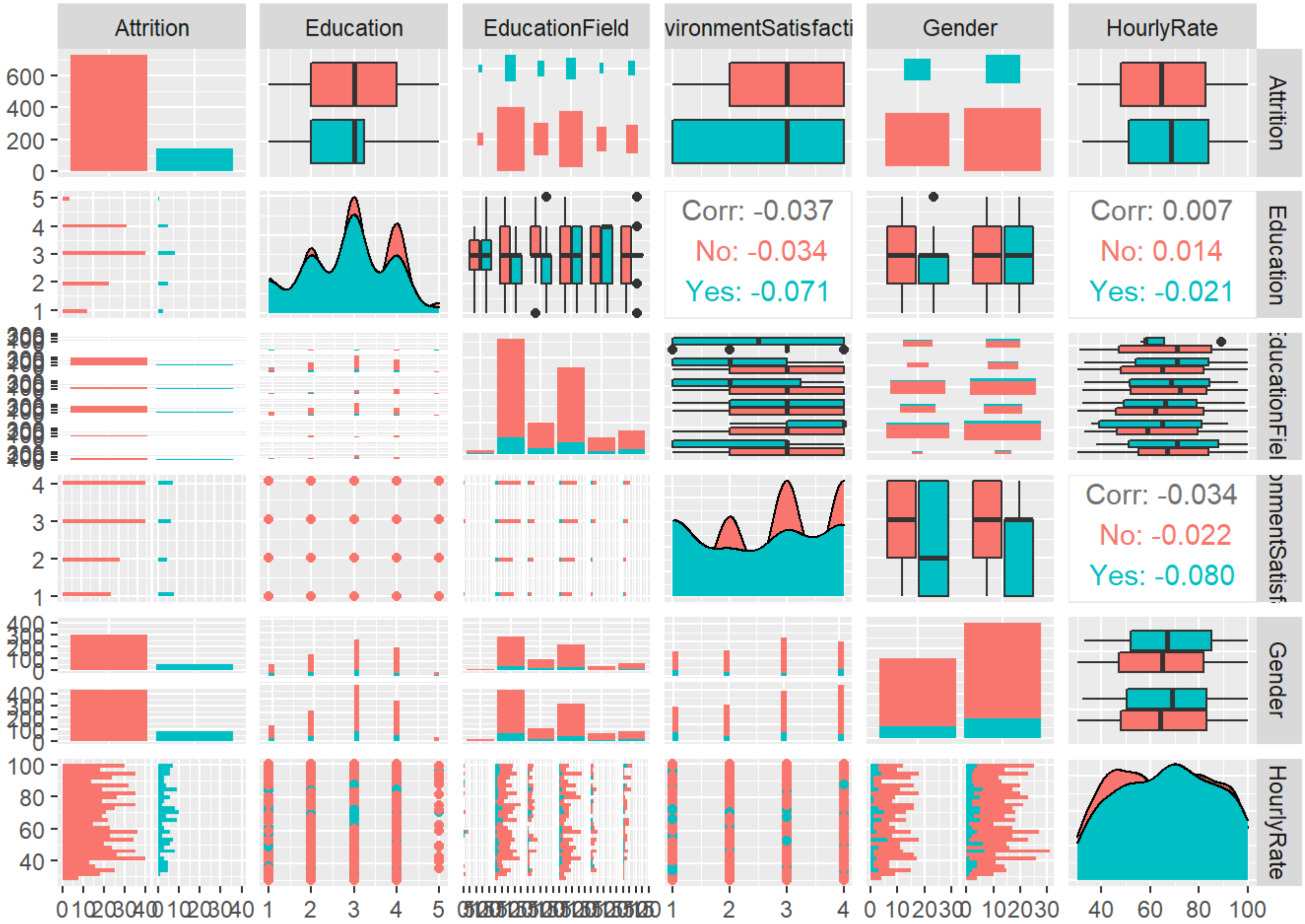
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



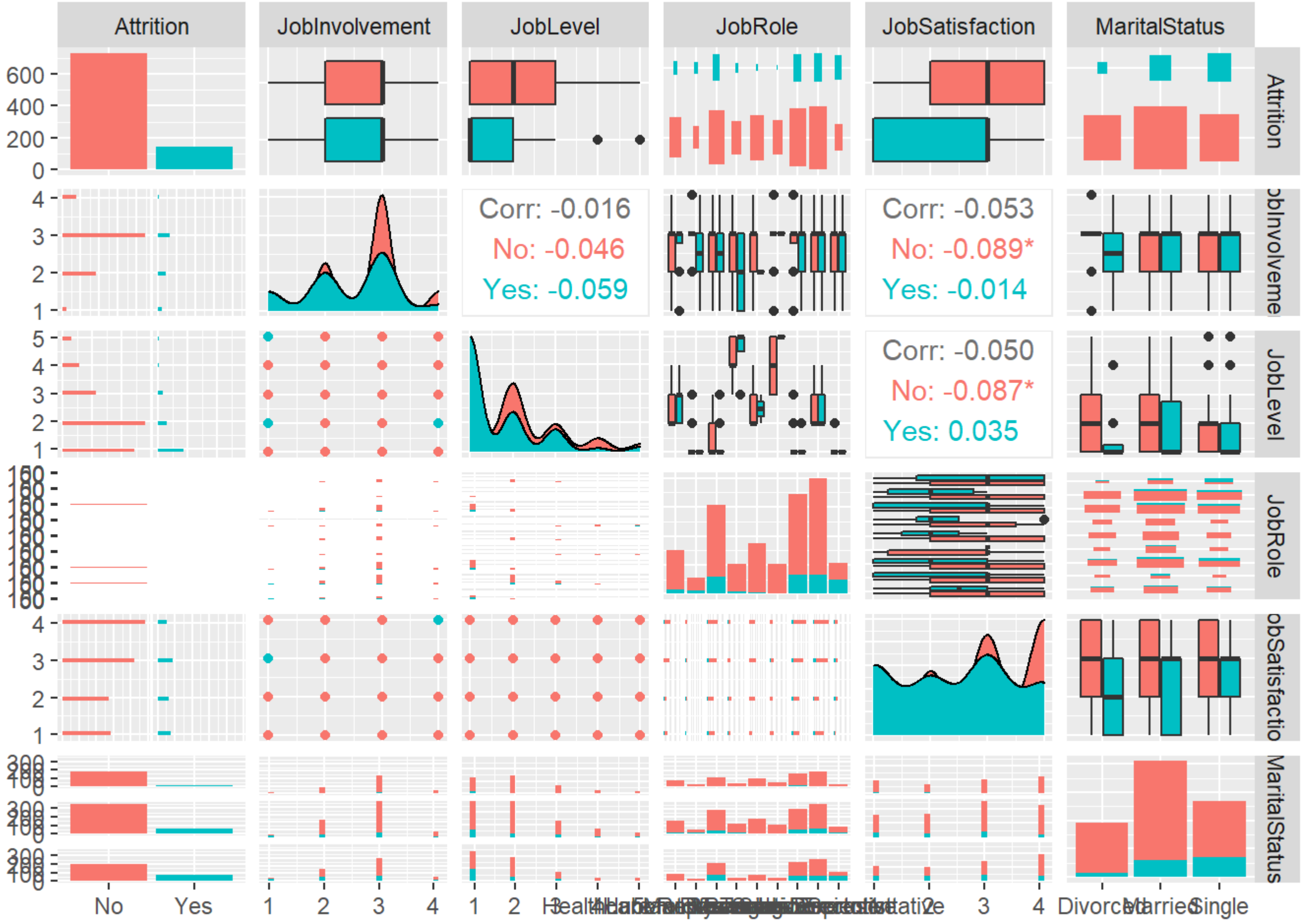
```
data %>%  select(Attrition,Education,EducationField,EnvironmentSatisfaction,Gender,HourlyRate) %>% ggpairs(aes(color = Attrition))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



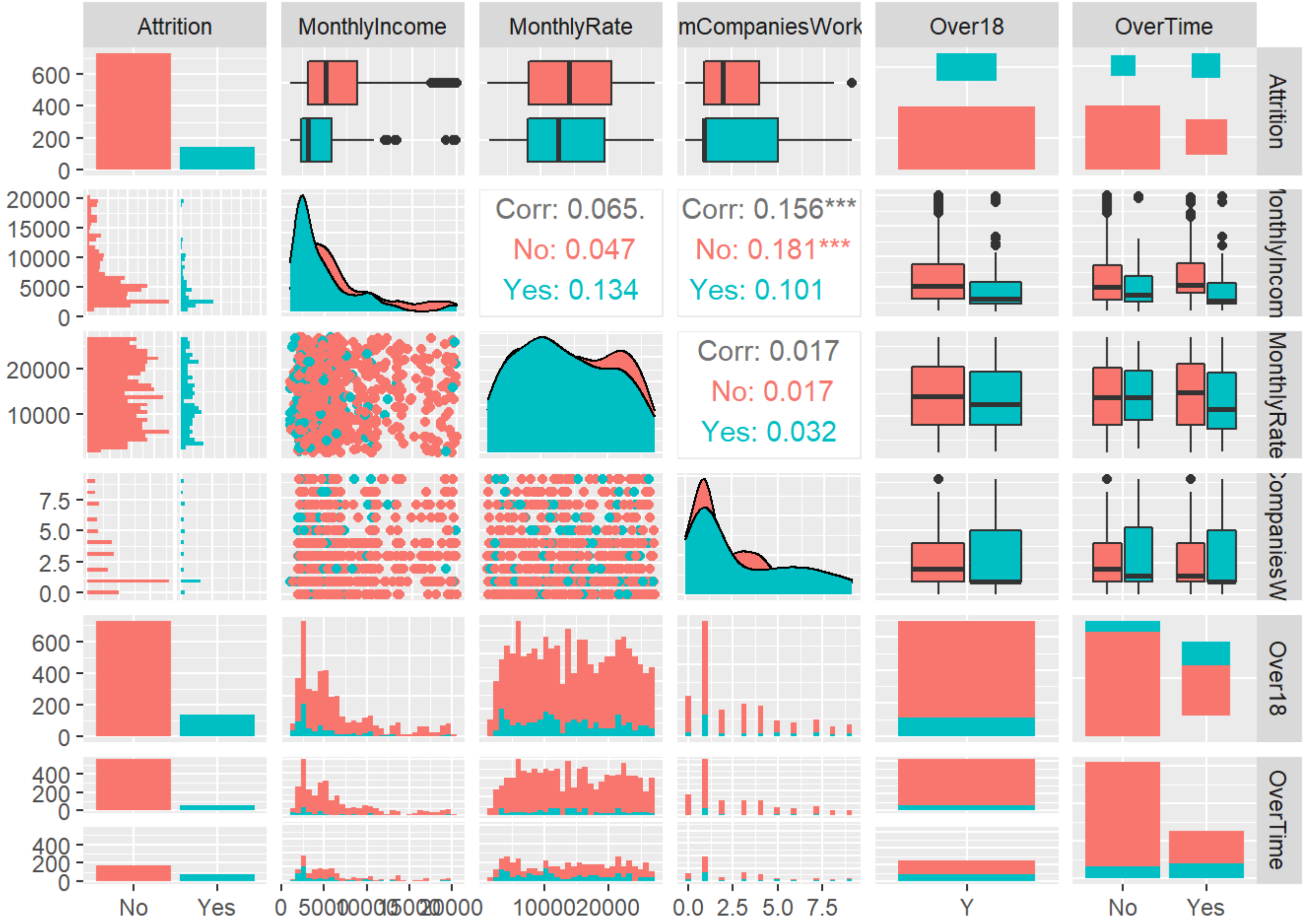
```
data %>%  select(Attrition,JobInvolvement,JobLevel,JobRole,JobSatisfaction,MaritalStatus) %>% ggpairs(aes(color = Attritio
n))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



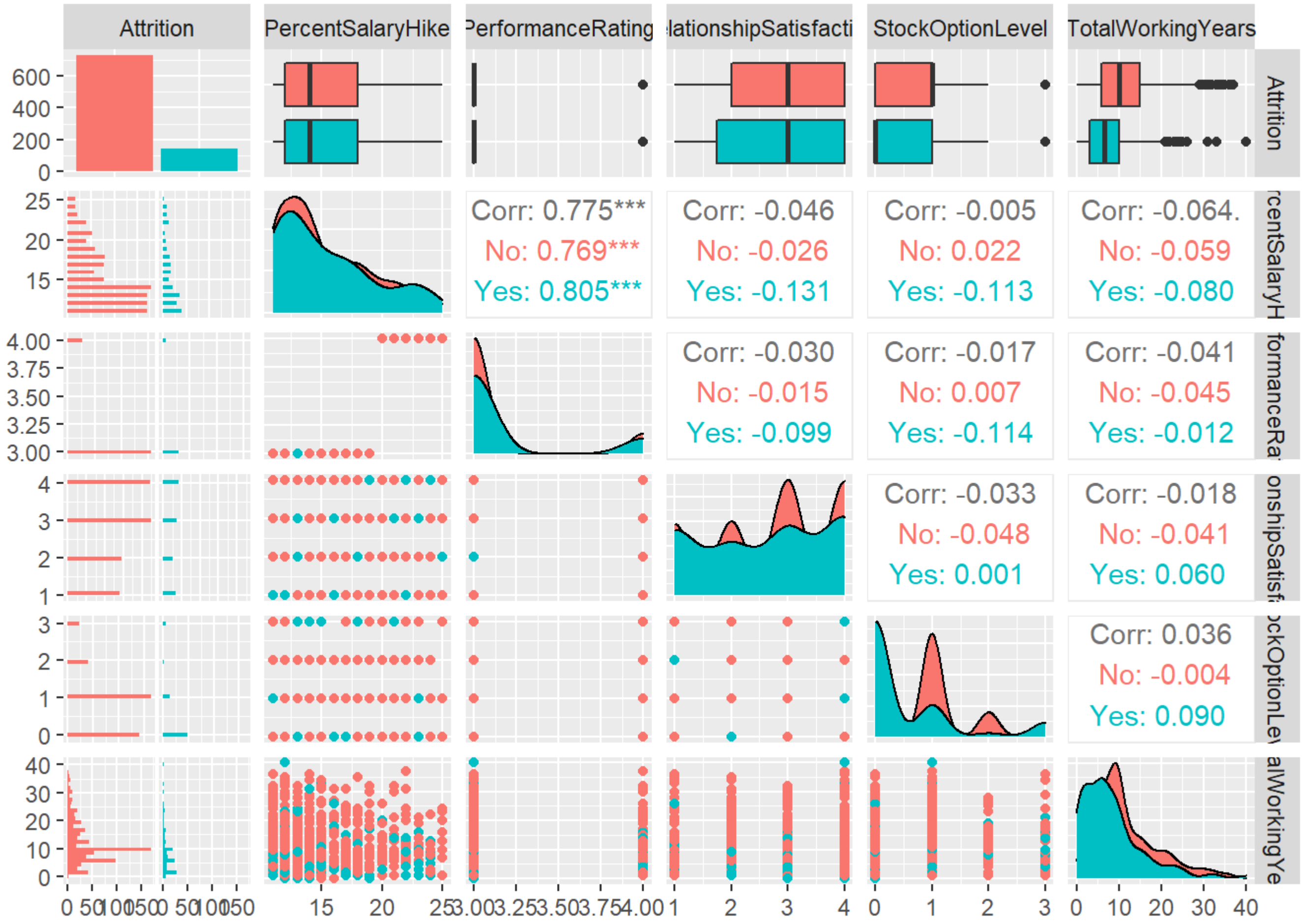
```
data %>%  select(Attrition,MonthlyIncome,MonthlyRate,NumCompaniesWorked,Over18,OverTime) %>% ggpairs(aes(color = Attrition))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



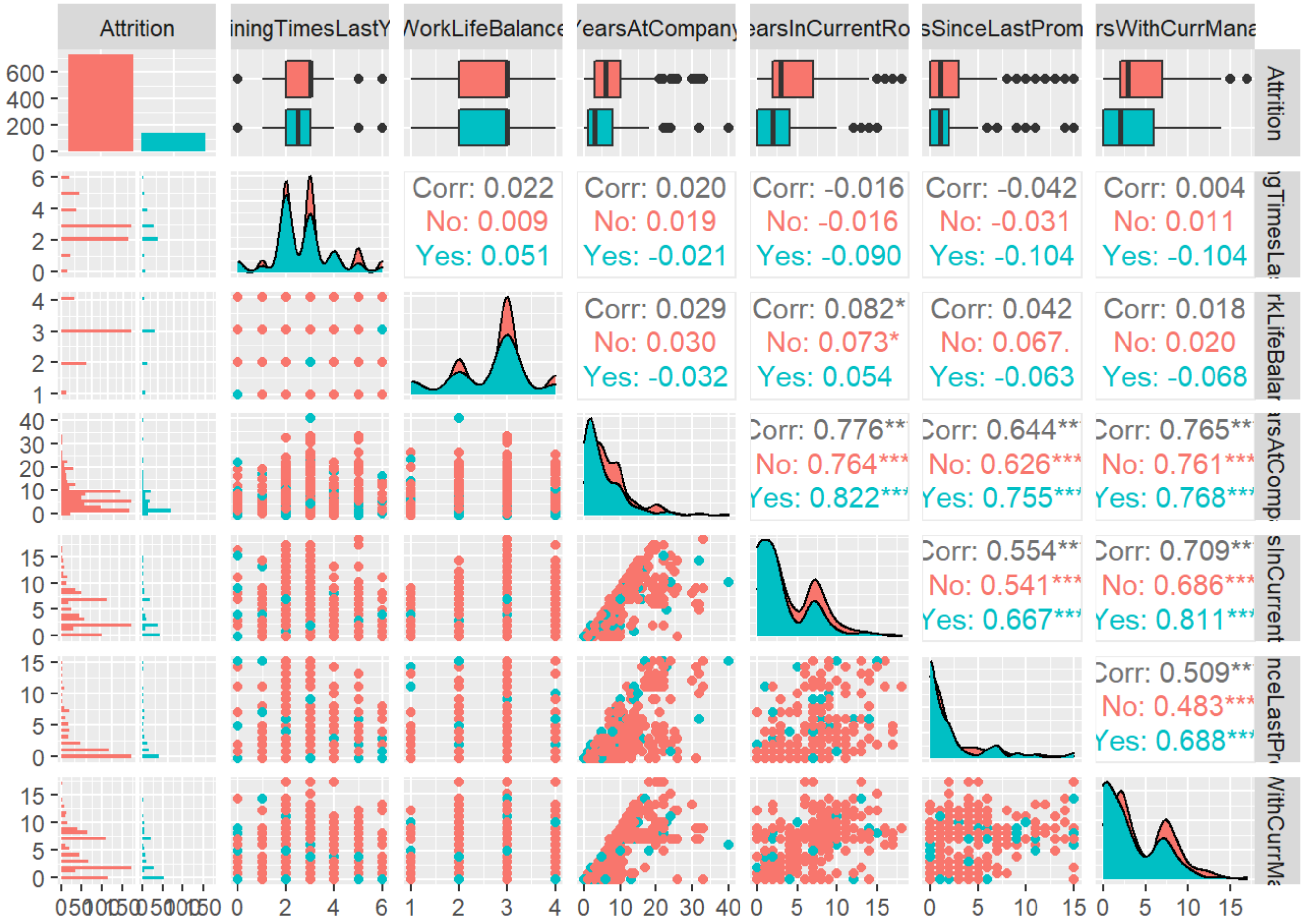
```
data %>%  select(Attrition,PercentSalaryHike,PerformanceRating,RelationshipSatisfaction,StockOptionLevel,TotalWorkingYears)
%>% ggpairs(aes(color = Attrition))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
data %>%  select(Attrition,TrainingTimesLastYear,WorkLifeBalance,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,Y
earsWithCurrManager) %>% ggpairs(aes(color = Attrition))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Attrition Analysis - Numeric

The same variables - Age, Distance from Home, Job Level, Monthly Income, Stock Option Level,Total Working Years, Years at Company, Years under Current Manager is highlighted numerically which validates our prediction in previous section.

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.2.2
```

```
newdata <- data %>% select(c(2,3:9,12:22,26:27,29:36))
dataatt <- newdata %>% group_by(Attrition)
skimr::skim(dataatt)
```

Data summary

Name	dataatt
Number of rows	870
Number of columns	29
________________________	
Column type frequency:	
factor	6
numeric	22
________________________	
Group variables	Attrition

Variable type: factor

skim_variable	Attrition	n_missing	complete_rate	ordered	n_unique	top_counts
BusinessTravel	No	0	1	FALSE	3	Tra: 524, Tra: 123, Non: 83
BusinessTravel	Yes	0	1	FALSE	3	Tra: 94, Tra: 35, Non: 11
Department	No	0	1	FALSE	3	Res: 487, Sal: 214, Hum: 29
Department	Yes	0	1	FALSE	3	Res: 75, Sal: 59, Hum: 6
EducationField	No	0	1	FALSE	6	Lif: 305, Med: 233, Mar: 80, Tec: 58
EducationField	Yes	0	1	FALSE	6	Lif: 53, Med: 37, Mar: 20, Tec: 17
Gender	No	0	1	FALSE	2	Mal: 429, Fem: 301
Gender	Yes	0	1	FALSE	2	Mal: 87, Fem: 53

skim_variable	Attrition	n_missing	complete_rate	ordered	n_unique	top_counts
JobRole	No	0	1	FALSE	9	Sal: 167, Res: 140, Lab: 123, Man: 85
JobRole	Yes	0	1	FALSE	9	Sal: 33, Res: 32, Lab: 30, Sal: 24
MaritalStatus	No	0	1	FALSE	3	Mar: 352, Sin: 199, Div: 179
MaritalStatus	Yes	0	1	FALSE	3	Sin: 70, Mar: 58, Div: 12

Variable type: numeric

skim_variable	Attrition	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	No	0	1	37.41	8.67	18	31.00	36.0	43.00	60	▂▇▇▃▂
Age	Yes	0	1	33.79	9.61	18	28.00	32.0	39.00	58	▃▇▃▂▂
DailyRate	No	0	1	821.16	401.41	111	483.75	828.5	1178.25	1499	▇▇▇▇▇
DailyRate	Yes	0	1	784.29	399.56	103	428.75	751.0	1110.75	1496	▇▇▆▆▆
DistanceFromHome	No	0	1	9.03	7.98	1	2.00	7.0	13.00	29	▇▅▁▂▂
DistanceFromHome	Yes	0	1	10.96	8.75	1	3.00	9.0	19.00	29	▇▅▂▂▃
Education	No	0	1	2.92	1.02	1	2.00	3.0	4.00	5	▂▅▇▆▁
Education	Yes	0	1	2.79	1.01	1	2.00	3.0	3.25	5	▂▅▇▅▁
EnvironmentSatisfaction	No	0	1	2.74	1.08	1	2.00	3.0	4.00	4	▅▆▁▇▇
EnvironmentSatisfaction	Yes	0	1	2.51	1.19	1	1.00	3.0	4.00	4	▇▅▁▇▇
HourlyRate	No	0	1	65.29	20.20	30	48.00	64.5	82.75	100	▇▇▆▇▇
HourlyRate	Yes	0	1	67.29	19.71	32	51.00	68.5	84.00	100	▆▇▇▇▇
JobInvolvement	No	0	1	2.78	0.67	1	2.00	3.0	3.00	4	▁▃▁▇▁
JobInvolvement	Yes	0	1	2.42	0.81	1	2.00	3.0	3.00	4	▂▅▁▇▁
JobLevel	No	0	1	2.12	1.09	1	1.00	2.0	3.00	5	▇▇▃▂▁
JobLevel	Yes	0	1	1.64	0.98	1	1.00	1.0	2.00	5	▇▃▂▁▁

skim_variable	Attrition	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
JobSatisfaction	No	0	1	2.76	1.11	1	2.00	3.0	4.00	4	▅▅▁▇▇
JobSatisfaction	Yes	0	1	2.44	1.09	1	1.00	3.0	3.00	4	▇▆▁▇▅
MonthlyIncome	No	0	1	6702.00	4675.47	1129	3162.00	5208.5	8736.50	19999	▇▅▂▁▂
MonthlyIncome	Yes	0	1	4764.79	3786.39	1081	2341.50	3171.0	5838.75	19859	▇▂▁▁▁
MonthlyRate	No	0	1	14460.12	7126.98	2094	8191.25	14235.5	20644.75	26997	▇▇▇▇▇
MonthlyRate	Yes	0	1	13624.29	6993.82	2396	8054.25	12651.0	19498.00	26959	▇▇▆▆▅
NumCompaniesWorked	No	0	1	2.66	2.47	0	1.00	2.0	4.00	9	▇▃▂▂▁
NumCompaniesWorked	Yes	0	1	3.08	2.77	0	1.00	1.0	5.00	9	▇▂▂▂▂
PerformanceRating	No	0	1	3.15	0.36	3	3.00	3.0	3.00	4	▇▁▁▁▂
PerformanceRating	Yes	0	1	3.16	0.37	3	3.00	3.0	3.00	4	▇▁▁▁▂
RelationshipSatisfaction	No	0	1	2.73	1.09	1	2.00	3.0	4.00	4	▅▅▁▇▇
RelationshipSatisfaction	Yes	0	1	2.61	1.16	1	1.75	3.0	4.00	4	▇▅▁▇▇
StockOptionLevel	No	0	1	0.84	0.84	0	0.00	1.0	1.00	3	▇▇▁▂▁
StockOptionLevel	Yes	0	1	0.49	0.90	0	0.00	0.0	1.00	3	▇▂▁▁▁
TotalWorkingYears	No	0	1	11.60	7.46	0	6.00	10.0	15.00	37	▆▇▃▁▁
TotalWorkingYears	Yes	0	1	8.19	7.16	0	3.00	6.5	10.00	40	▇▃▁▁▁
TrainingTimesLastYear	No	0	1	2.87	1.28	0	2.00	3.0	3.00	6	▂▇▇▂▃
TrainingTimesLastYear	Yes	0	1	2.65	1.23	0	2.00	2.5	3.00	6	▂▇▆▂▂
WorkLifeBalance	No	0	1	2.81	0.69	1	2.00	3.0	3.00	4	▁▃▁▇▂
WorkLifeBalance	Yes	0	1	2.64	0.82	1	2.00	3.0	3.00	4	▂▃▁▇▁
YearsAtCompany	No	0	1	7.30	5.94	0	3.00	6.0	10.00	33	▇▅▁▁▁
YearsAtCompany	Yes	0	1	5.19	6.17	0	1.00	3.0	8.00	40	▇▂▁▁▁

skim_variable	Attrition	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
YearsInCurrentRole	No	0	1	4.45	3.64	0	2.00	3.0	7.00	18	▇▅▂▁▁
YearsInCurrentRole	Yes	0	1	2.91	3.33	0	0.00	2.0	4.00	15	▇▁▂▁▁
YearsSinceLastPromotion	No	0	1	2.18	3.15	0	0.00	1.0	3.00	15	▇▁▁▁▁
YearsSinceLastPromotion	Yes	0	1	2.14	3.40	0	0.00	1.0	2.00	15	▇▁▁▁▁
YearsWithCurrManager	No	0	1	4.37	3.59	0	2.00	3.0	7.00	17	▇▂▅▁▁
YearsWithCurrManager	Yes	0	1	2.94	3.24	0	0.00	2.0	6.00	14	▇▂▂▁▁

Top 3 Attrition Reason

From the Attrition Analysis in previous 2 sections - Age, Business Travel,Distance from Home, Job Level, Monthly Income, Stock Option Level,Total Working Years, Years at Company were identified as important inputs for Attrition.I ran numerous models with knn and NB selecting the 3 variables at a time. The best model that I got was using Naive Bayes model with inputs Age, Business Travel , and Work Year(which was changed to Factor). The data was skewed heavily towards “No” attrition which meant that random sampling of total dataset didn’t yield enough “Yes” attrition. To tackle this issue, dataset was filtered into “Yes” and “No” attrition and ~80 % of “Yes” were samples every time along with ~75 % of “No”. 50 seeds were taken to get the mean for accuracy, sensitivity and Specificity.

```
library(caret)
```

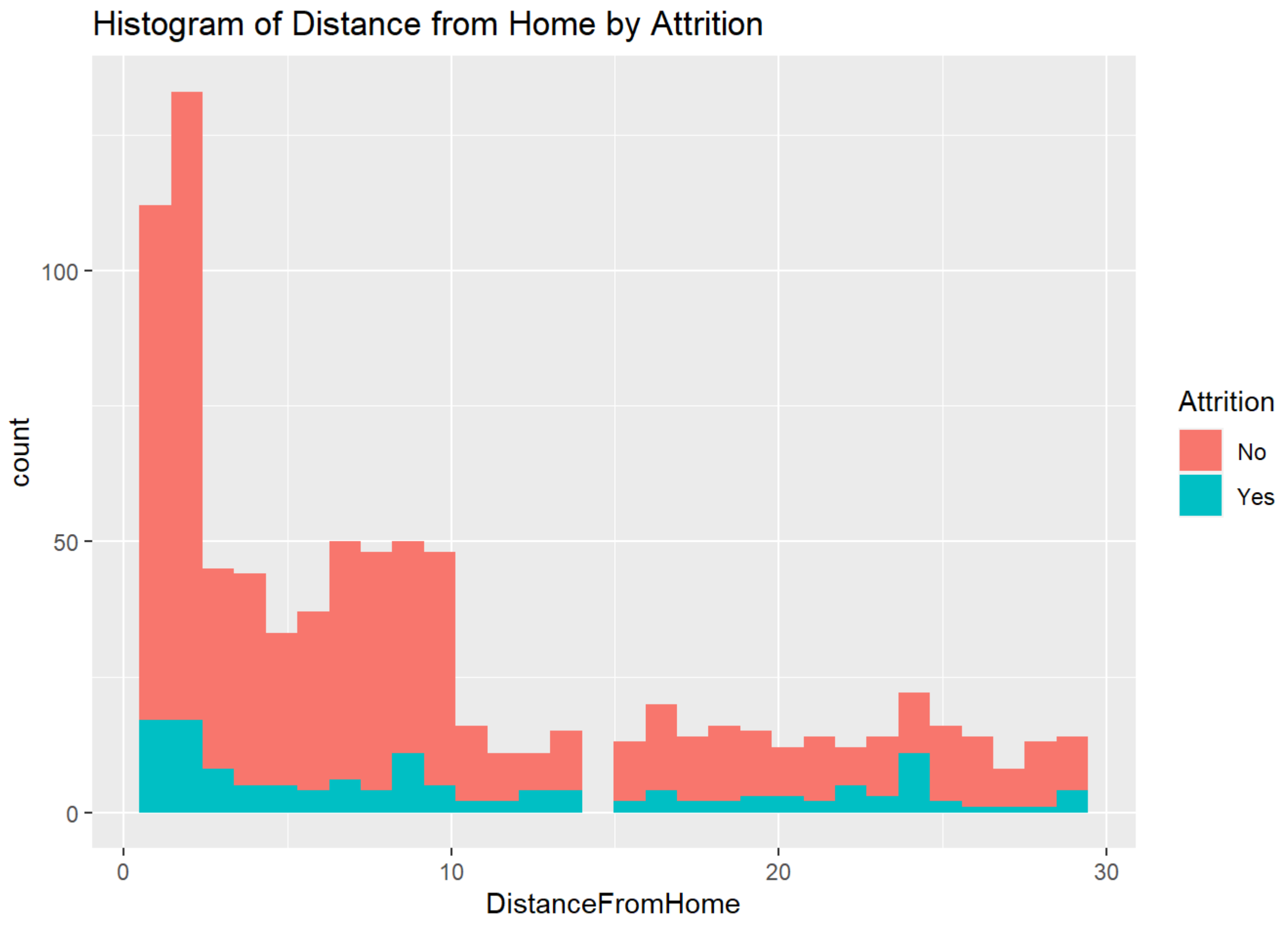
```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

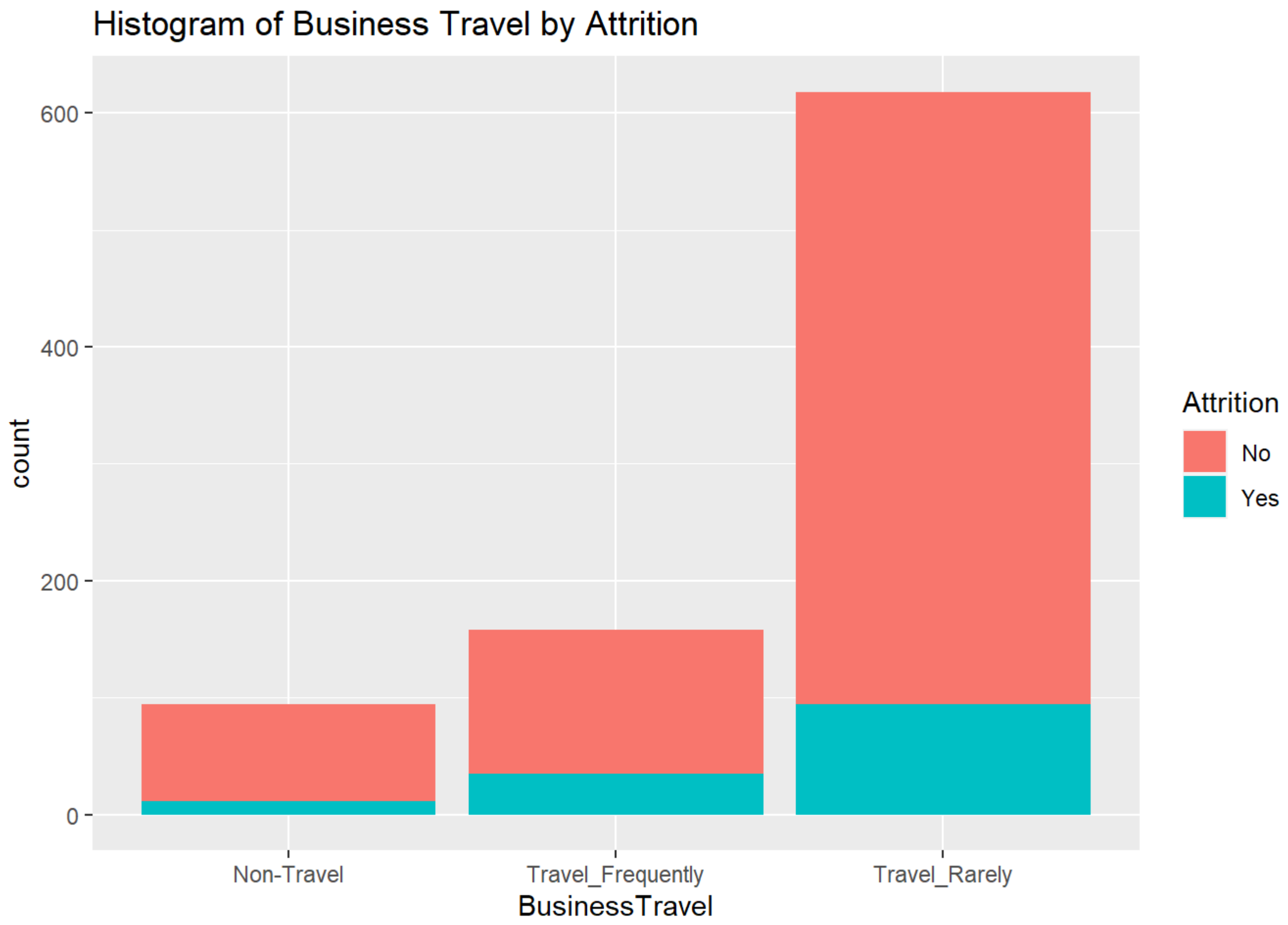
```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library (e1071)
# Histograms for key inputs
datanb <- data
datanb %>% ggplot(aes(x= DistanceFromHome, fill=Attrition)) + geom_histogram() + ggtitle("Histogram of Distance from Home by Attrition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

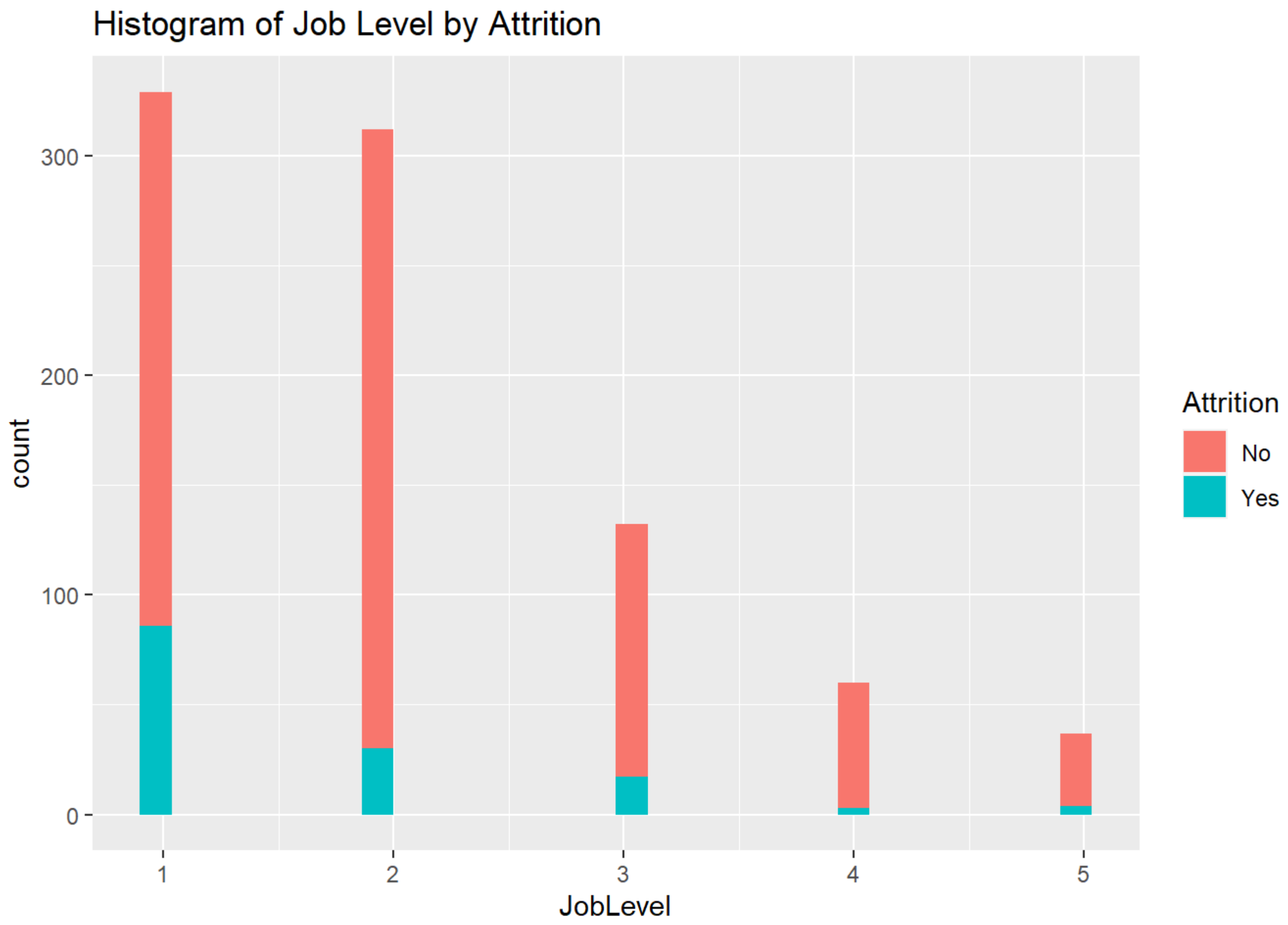


```
datanb %>% ggplot(aes(x= BusinessTravel, fill=Attrition)) + geom_bar(stat="count") + ggtitle("Histogram of Business Travel by Attrition")
```



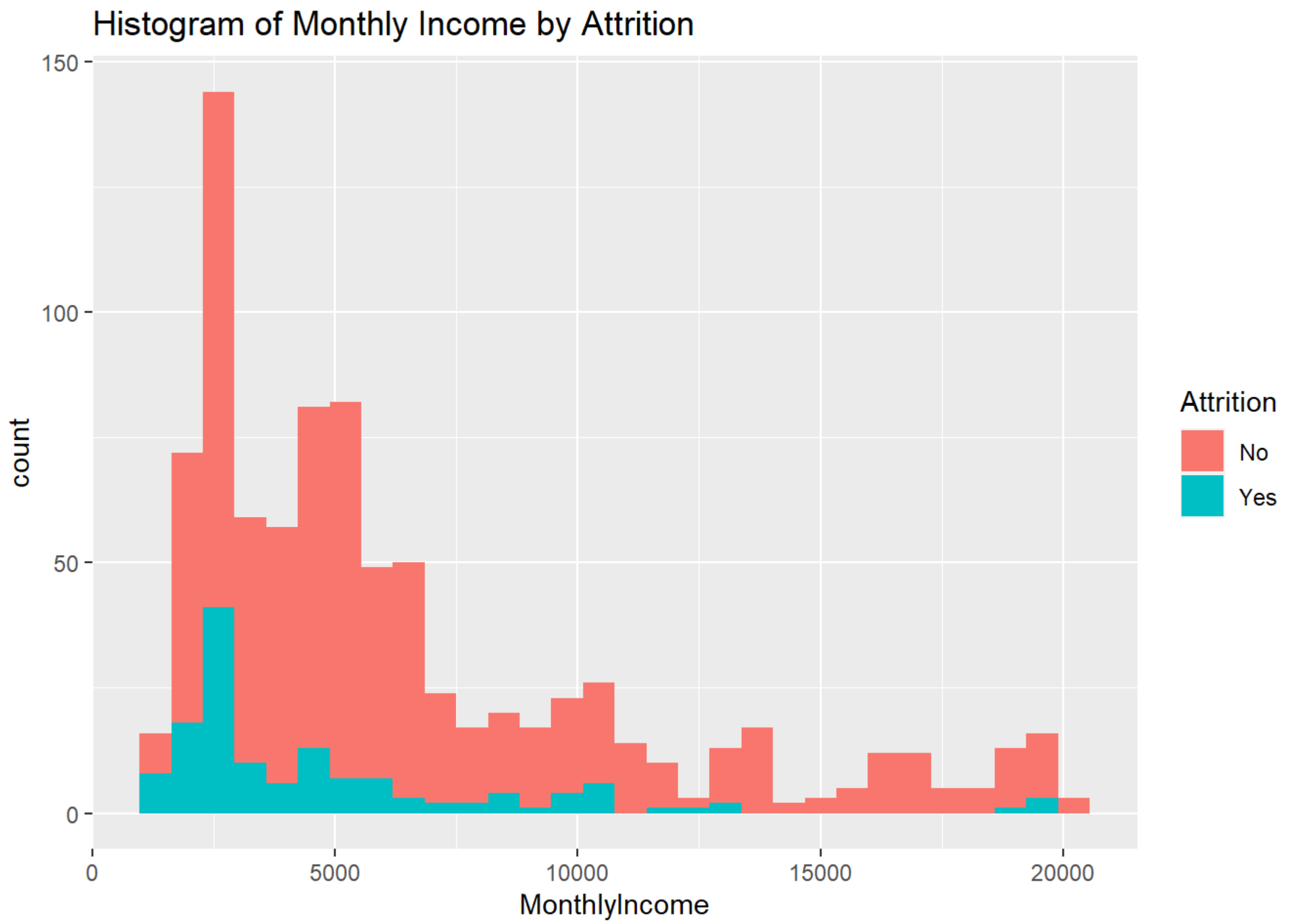
```
datanb %>% ggplot(aes(x= JobLevel, fill=Attrition)) + geom_histogram() + ggtitle("Histogram of Job Level by Attrition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



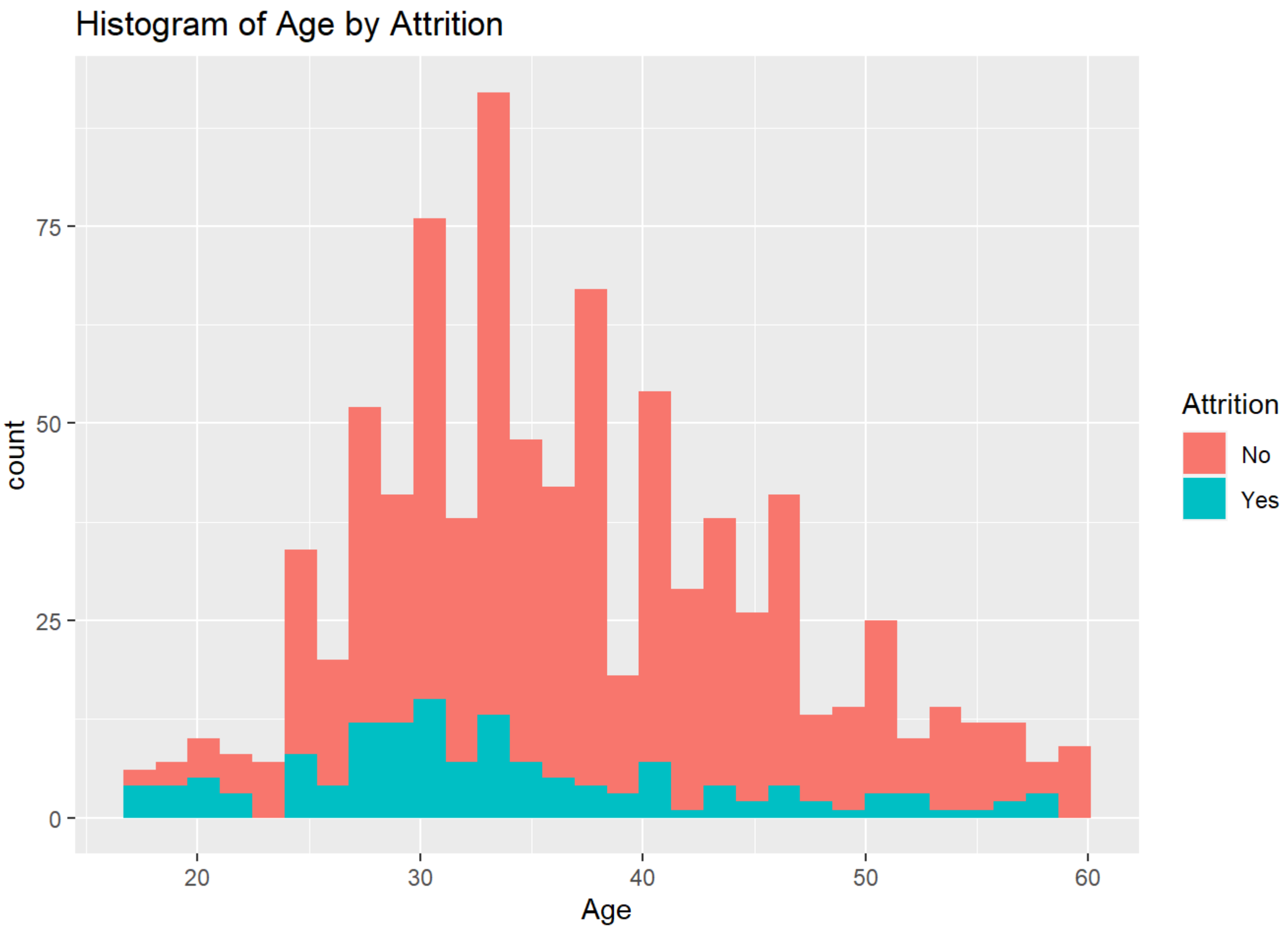
```
datanb %>% ggplot(aes(x= MonthlyIncome, fill=Attrition)) + geom_histogram() + ggtitle("Histogram of Monthly Income by Attrit
ion")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



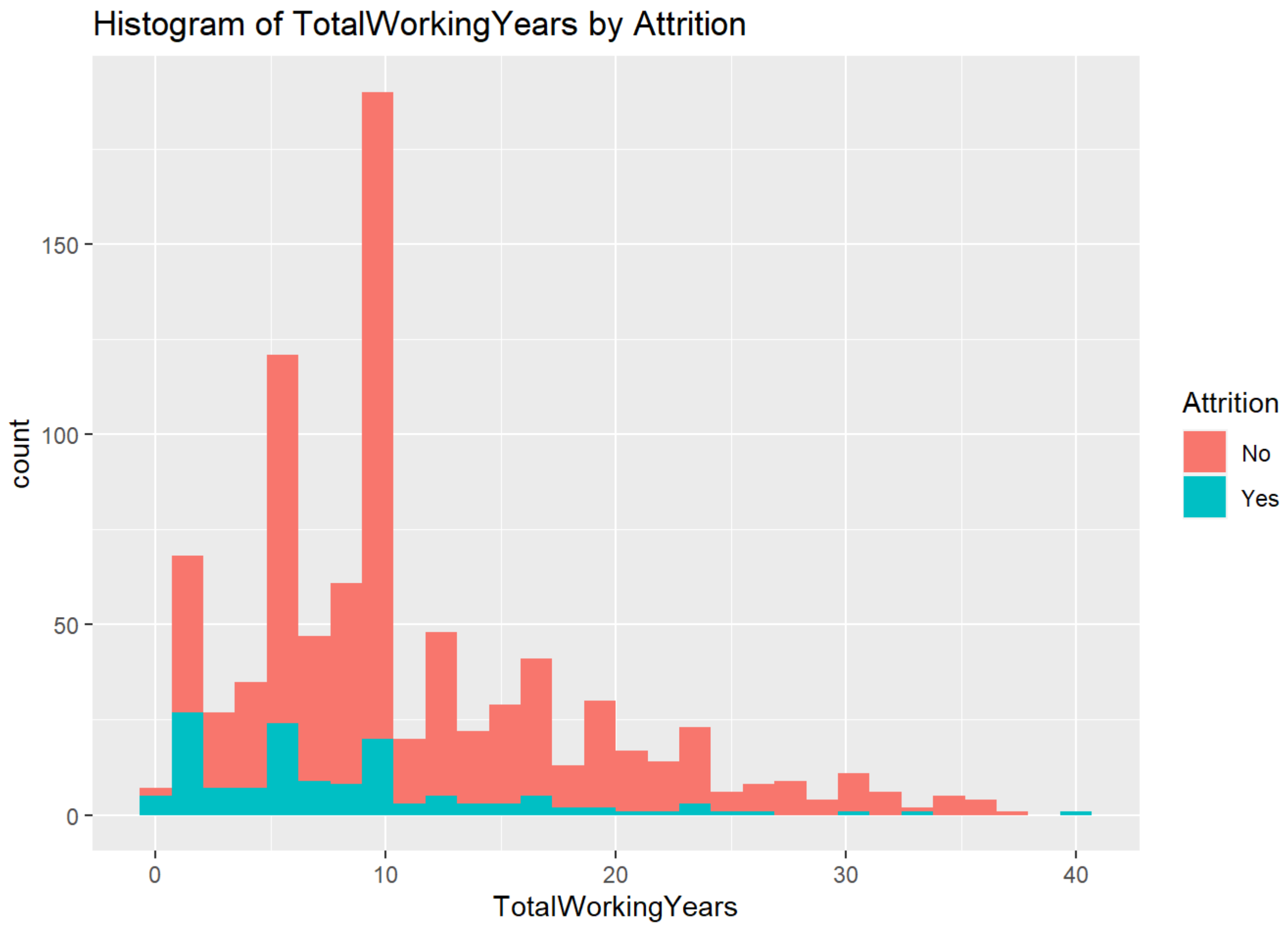
```
datanb %>% ggplot(aes(x= Age, fill=Attrition)) + geom_histogram() + ggtitle("Histogram of Age by Attrition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



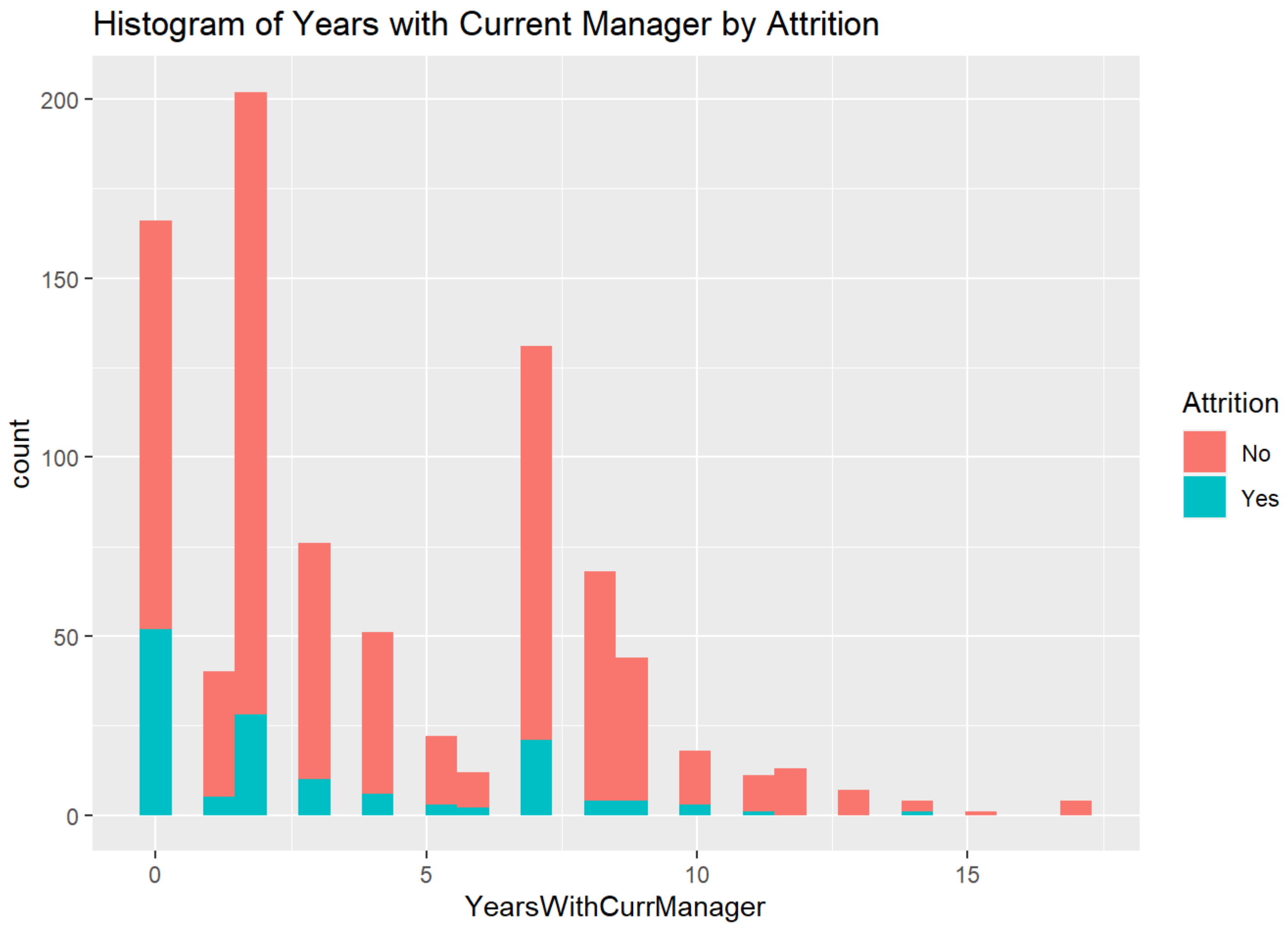
```
datanb %>% ggplot(aes(x= TotalWorkingYears , fill=Attrition)) + geom_histogram() + ggtitle("Histogram of TotalWorkingYears b
y Attrition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
datanb %>% ggplot(aes(x= YearsWithCurrManager  , fill=Attrition)) + geom_histogram() + ggtitle("Histogram of Years with Curr
ent Manager by Attrition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Histogram of Years with Current Manager by Attrition
count
200
150
100
50
0
0
5
10
15
YearsWithCurrManager
Attrition
No
Yes

```
# Changing Work years to Factor with Levels based on Histogram
datanb$WorkYearFactor = cut(datanb$TotalWorkingYears, breaks = c(-1,11,21,40),labels = c("1","2","3"))

# Changing Age to Factor with Levels based on Histogram
datanb$AgeFactor = cut(datanb$Age, breaks = c(17,25,30,35,40,45,61),labels = c("1", "2", "3", "4", "5","6"))

# Changing Monthly Income to Factor with Levels based on Histogram
datanb$SalaryFactor = cut(datanb$MonthlyIncome, breaks = c(1080,3000,6000,12000,25000),labels = c("<3k","3k to 6k","6k to 12k",">12k"))

# Changing Years with Current Manager to Factor with Levels based on Histogram
datanb$YearsWithCurrManagerFactor  = cut(datanb$YearsWithCurrManager, breaks = c(-1,5,10,17),labels = c("Low","Med","High"))

# Creating dataset with "Yes" and "No" Attrition
datayes <- datanb %>% filter(Attrition =="Yes")
datano <- datanb %>% filter(Attrition =="No")

summary(datanb)
```

```
##        ID             Age        Attrition           BusinessTravel
##  Min.   :  1.0   Min.   :18.00   No :730   Non-Travel       : 94
##  1st Qu.:218.2   1st Qu.:30.00   Yes:140   Travel_Frequently:158
##  Median :435.5   Median :35.00             Travel_Rarely    :618
##  Mean   :435.5   Mean   :36.83
##  3rd Qu.:652.8   3rd Qu.:43.00
##  Max.   :870.0   Max.   :60.00
##
##    DailyRate                       Department  DistanceFromHome   Education
##  Min.   : 103.0   Human Resources       : 35   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 472.5   Research & Development:562   1st Qu.: 2.000   1st Qu.:2.000
##  Median : 817.5   Sales                 :273   Median : 7.000   Median :3.000
##  Mean   : 815.2                                Mean   : 9.339   Mean   :2.901
##  3rd Qu.:1165.8                                3rd Qu.:14.000   3rd Qu.:4.000
##  Max.   :1499.0                                Max.   :29.000   Max.   :5.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 15    Min.   :1     Min.   :   1.0   Min.   :1.000
##  Life Sciences   :358    1st Qu.:1     1st Qu.: 477.2   1st Qu.:2.000
##  Marketing       :100    Median :1     Median :1039.0   Median :3.000
##  Medical         :270    Mean   :1     Mean   :1029.8   Mean   :2.701
##  Other           : 52    3rd Qu.:1     3rd Qu.:1561.5   3rd Qu.:4.000
##  Technical Degree: 75    Max.   :1     Max.   :2064.0   Max.   :4.000
##
##     Gender      HourlyRate     JobInvolvement     JobLevel
##  Female:354   Min.   : 30.00   Min.   :1.000   Min.   :1.000
##  Male  :516   1st Qu.: 48.00   1st Qu.:2.000   1st Qu.:1.000
##               Median : 66.00   Median :3.000   Median :2.000
##               Mean   : 65.61   Mean   :2.723   Mean   :2.039
##               3rd Qu.: 83.00   3rd Qu.:3.000   3rd Qu.:3.000
##               Max.   :100.00   Max.   :4.000   Max.   :5.000
##
##                        JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
##  Sales Executive          :200   Min.   :1.000   Divorced:191   Min.   : 1081
##  Research Scientist       :172   1st Qu.:2.000   Married :410   1st Qu.: 2840
##  Laboratory Technician    :153   Median :3.000   Single  :269   Median : 4946
##  Manufacturing Director   : 87   Mean   :2.709                  Mean   : 6390
##  Healthcare Representative: 76   3rd Qu.:4.000                  3rd Qu.: 8182
##  Sales Representative     : 53   Max.   :4.000                  Max.   :19999
##  (Other)                  :129
##   MonthlyRate    NumCompaniesWorked Over18  OverTime  PercentSalaryHike
##  Min.   : 2094   Min.   :0.000      Y:870   No :618   Min.   :11.0
##  1st Qu.: 8092   1st Qu.:1.000              Yes:252   1st Qu.:12.0
##  Median :14074   Median :2.000                        Median :14.0
##  Mean   :14326   Mean   :2.728                        Mean   :15.2
##  3rd Qu.:20456   3rd Qu.:4.000                        3rd Qu.:18.0
##  Max.   :26997   Max.   :9.000                        Max.   :25.0
##
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.000     Median :3.000            Median :80    Median :1.0000
##  Mean   :3.152     Mean   :2.707            Mean   :80    Mean   :0.7839
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000
##
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000
##  Mean   :11.05     Mean   :2.832         Mean   :2.782   Mean   : 6.962
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.:10.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
##
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager WorkYearFactor
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.00        1:576
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.00        2:200
##  Median : 3.000     Median : 1.000          Median : 3.00        3: 94
##  Mean   : 4.205     Mean   : 2.169          Mean   : 4.14
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.00
##  Max.   :18.000     Max.   :15.000          Max.   :17.00
##
##  AgeFactor    SalaryFactor YearsWithCurrManagerFactor
##  1: 72     <3k      :242   Low :557
##  2:150     3k to 6k :302   Med :273
##  3:217     6k to 12k:215   High: 40
##  4:156     >12k     :111
##  5:118
##  6:157
##
```

```
# NaiveBayes Model with Age(2),Business Travel(4) and Work Year Factor (37)
AccHolder = numeric(50)
SensHolder = numeric(50)
SpecHolder = numeric(50)

for (seed in 1:50)
{
set.seed(seed)
trainIndices_yes = sample(seq(1:140),115)
trainIndices_no = sample(seq(1:730),555)
trainAttrition = rbind(datayes[trainIndices_yes,] , datano[trainIndices_no,])
testAttrition = rbind(datayes[-trainIndices_yes,], datano[-trainIndices_no,])
model = naiveBayes(trainAttrition[,c(2,4,37)],trainAttrition$Attrition)
CM = confusionMatrix(table(testAttrition$Attrition, predict(model,testAttrition[,c(2,4,37)])))
AccHolder[seed] = CM$overall[1]
SensHolder[seed] = CM$byClass[1]
SpecHolder[seed] = CM$byClass[2]
}

mean(AccHolder) # Mean Accuracy = 0.88
```

```
## [1] 0.8761
```

```
#Standard Error of the Mean
sd(AccHolder)/sqrt(50)
```

```
## [1] 0.0002958902
```

```
mean(SensHolder) # Mean Sensitivity = 0.88
```

```
## [1] 0.8759673
```

```
#Standard Error of the Mean
sd(SensHolder)/sqrt(50)
```

```
## [1] 0.000260205
```

```
mean(SpecHolder,na.rm = TRUE) # Mean Specificity = 1
```

```
## [1] 1
```

```
#Standard Error of the Mean
sd(SensHolder)/sqrt(50)
```

```
## [1] 0.000260205
```

Best Model to Predict Attrition

I realize that due to skewness, we need a model with high enough accuracy, sensitivity, but not too high specificity. I used Naive Bayes and knn models. The best values that I got was using Naive Bayes. I used 80% for Traning set and 20 % for test set. Due to skewness, the sets were bound from seperate Attrition and No Attrition datasets. The factors selected for my models are Age, Business Travel, Monthly Income, Work Years in factor and Years with Current Manager in Factor.

```
# NaiveBayes Model with Age(2),Business Travel(4),Monthly Income (20),Work Year Factor (37), and Years with Current Manager Factor (40) to check if the model is stable.
AccHolder = numeric(50)
SensHolder = numeric(50)
SpecHolder = numeric(50)

for (seed in 1:50)
{
set.seed(seed)
trainIndices_yes = sample(seq(1:140),112)
trainIndices_no = sample(seq(1:730),584)
trainAttrition = rbind(datayes[trainIndices_yes,] , datano[trainIndices_no,])
testAttrition = rbind(datayes[-trainIndices_yes,], datano[-trainIndices_no,])
model = naiveBayes(trainAttrition[,c(2,4,20,40,37)],trainAttrition$Attrition)
CM = confusionMatrix(table(testAttrition$Attrition, predict(model,testAttrition[,c(2,4,20,40,37)])))
AccHolder[seed] = CM$overall[1]
SensHolder[seed] = CM$byClass[1]
SpecHolder[seed] = CM$byClass[2]
}

mean(AccHolder) # Mean Accuracy
```

```
## [1] 0.8462069
```

```
mean(SensHolder) # Mean Sensitivity
```

```
## [1] 0.8494159
```

```
mean(SpecHolder,na.rm = TRUE) # Mean Specificity
```

```
## [1] 0.7765306
```

```
AccHolder
```

```
##  [1] 0.8333333 0.8563218 0.8390805 0.8620690 0.8275862 0.8448276 0.8448276
##  [8] 0.8563218 0.8505747 0.8505747 0.8505747 0.8448276 0.8390805 0.8448276
## [15] 0.8333333 0.8563218 0.8563218 0.8563218 0.8390805 0.8448276 0.8563218
## [22] 0.8448276 0.8505747 0.8505747 0.8505747 0.7931034 0.8275862 0.8563218
## [29] 0.8333333 0.8448276 0.8390805 0.8563218 0.8390805 0.8390805 0.8563218
## [36] 0.8448276 0.8505747 0.8563218 0.8505747 0.8448276 0.8505747 0.8505747
## [43] 0.8505747 0.8448276 0.8448276 0.8620690 0.8448276 0.8505747 0.8390805
## [50] 0.8563218
```

```
SensHolder
```

```
##  [1] 0.8461538 0.8538012 0.8430233 0.8588235 0.8452381 0.8562874 0.8439306
##  [8] 0.8579882 0.8571429 0.8488372 0.8571429 0.8479532 0.8390805 0.8439306
## [15] 0.8502994 0.8538012 0.8538012 0.8579882 0.8470588 0.8439306 0.8538012
## [22] 0.8479532 0.8529412 0.8488372 0.8488372 0.8313253 0.8536585 0.8538012
## [29] 0.8421053 0.8439306 0.8430233 0.8579882 0.8430233 0.8470588 0.8538012
## [36] 0.8479532 0.8488372 0.8538012 0.8488372 0.8520710 0.8529412 0.8529412
## [43] 0.8488372 0.8439306 0.8439306 0.8588235 0.8439306 0.8488372 0.8430233
## [50] 0.8538012
```

```
SpecHolder
```

```
##  [1] 0.4000000 1.0000000 0.5000000 1.0000000 0.3333333 0.5714286 1.0000000
##  [8] 0.8000000 0.6666667 1.0000000 0.6666667 0.6666667        NA 1.0000000
## [15] 0.4285714 1.0000000 1.0000000 0.8000000 0.5000000 1.0000000 1.0000000
## [22] 0.6666667 0.7500000 1.0000000 1.0000000 0.0000000 0.4000000 1.0000000
## [29] 0.3333333 1.0000000 0.5000000 0.8000000 0.5000000 0.5000000 1.0000000
## [36] 0.6666667 1.0000000 1.0000000 1.0000000 0.6000000 0.7500000 0.7500000
## [43] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 0.5000000
## [50] 1.0000000
```

```
# The mean values prove that the model is stable.

# The best seed that gave me a model high for accuracy and specifity and not too high for specifity is seed(9)
# Using seed 9 for the best model.
# Best prediction model
set.seed(9)
trainIndices_yes = sample(seq(1:140),112)
trainIndices_no = sample(seq(1:730),584)
trainAttrition = rbind(datayes[trainIndices_yes,] , datano[trainIndices_no,])
testAttrition = rbind(datayes[-trainIndices_yes,], datano[-trainIndices_no,])
model = naiveBayes(trainAttrition[,c(2,4,20,40,37)],trainAttrition$Attrition)
CM = confusionMatrix(table(testAttrition$Attrition, predict(model,testAttrition[,c(2,4,20,40,37)])))
model
```

```
## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = trainAttrition[, c(2, 4, 20, 40, 37)], 
##     y = trainAttrition$Attrition)
## 
## A-priori probabilities:
## trainAttrition$Attrition
##        No       Yes 
## 0.8390805 0.1609195 
## 
## Conditional probabilities:
##                         Age
## trainAttrition$Attrition     [,1]     [,2]
##                      No  37.52397 8.570729
##                      Yes 33.43750 9.032256
## 
##                         BusinessTravel
## trainAttrition$Attrition Non-Travel Travel_Frequently Travel_Rarely
##                      No  0.10787671        0.17123288    0.72089041
##                      Yes 0.05357143        0.25000000    0.69642857
## 
##                         MonthlyIncome
## trainAttrition$Attrition     [,1]     [,2]
##                      No  6804.099 4761.912
##                      Yes 4838.009 3906.460
## 
##                         YearsWithCurrManagerFactor
## trainAttrition$Attrition        Low        Med       High
##                      No  0.60787671 0.33732877 0.05479452
##                      Yes 0.72321429 0.25892857 0.01785714
## 
##                         WorkYearFactor
## trainAttrition$Attrition          1          2          3
##                      No  0.61643836 0.25856164 0.12500000
##                      Yes 0.77678571 0.15178571 0.07142857
```

```
CM
```

```
## Confusion Matrix and Statistics
## 
## 
##        No Yes
##   No  144   2
##   Yes  24   4
## 
##                Accuracy : 0.8506
##                  95% CI : (0.7888, 0.9)
##     No Information Rate : 0.9655
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : 0.1892
## 
##  Mcnemar's Test P-Value : 3.814e-05
## 
##             Sensitivity : 0.8571
##             Specificity : 0.6667
##          Pos Pred Value : 0.9863
##          Neg Pred Value : 0.1429
##              Prevalence : 0.9655
##          Detection Rate : 0.8276
##    Detection Prevalence : 0.8391
##       Balanced Accuracy : 0.7619
## 
##        'Positive' Class : No
## 
```

```
# Accuracy = 0.8506
# Sensitivity = 0.8571
# Specificity = 0.6667

# Data modification for No Attrition Dataset

head(Casestudy2NoA)
```

```
##     ID Age    BusinessTravel DailyRate             Department DistanceFromHome
## 1 1171  35     Travel_Rarely       750 Research & Development               28
## 2 1172  33     Travel_Rarely       147        Human Resources                2
## 3 1173  26     Travel_Rarely      1330 Research & Development               21
## 4 1174  55     Travel_Rarely      1311 Research & Development                2
## 5 1175  29     Travel_Rarely      1246                  Sales               19
## 6 1176  51 Travel_Frequently      1456 Research & Development                1
##   Education  EducationField EmployeeCount EmployeeNumber
## 1         3   Life Sciences             1           1596
## 2         3 Human Resources             1           1207
## 3         3         Medical             1           1107
## 4         3   Life Sciences             1            505
## 5         3   Life Sciences             1           1497
## 6         4         Medical             1            145
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2   Male         46              4        2
## 2                       2   Male         99              3        1
## 3                       1   Male         37              3        1
## 4                       3 Female         97              3        4
## 5                       3   Male         77              2        2
## 6                       1 Female         30              2        3
##                     JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1     Laboratory Technician               3       Married          3407
## 2           Human Resources               3       Married          3600
## 3     Laboratory Technician               3      Divorced          2377
## 4                   Manager               4        Single         16659
## 5           Sales Executive               3      Divorced          8620
## 6 Healthcare Representative               1        Single          7484
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       25348                  1      Y       No                17
## 2        8429                  1      Y       No                13
## 3       19373                  1      Y       No                20
## 4       23258                  2      Y      Yes                13
## 5       23757                  1      Y       No                14
## 6       25796                  3      Y       No                20
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                        4            80                2
## 2                 3                        4            80                1
## 3                 4                        3            80                1
## 4                 3                        3            80                0
## 5                 3                        3            80                2
## 6                 4                        3            80                0
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                10                     3               2             10
## 2                 5                     2               3              5
## 3                 1                     0               2              1
## 4                30                     2               3              5
## 5                10                     3               3             10
## 6                23                     1               2             13
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  9                       6                    8
## 2                  4                       1                    4
## 3                  1                       0                    0
## 4                  4                       1                    2
## 5                  7                       0                    4
## 6                 12                      12                    8
```

```
Casestudy2NoA$SalaryFactor = cut(Casestudy2NoA$MonthlyIncome, breaks = c(1080,3000,6000,12000,25000),labels = c("<3k","3k to
6k","6k to 12k",">12k"))
Casestudy2NoA$BusinessTravel = as.factor(Casestudy2NoA$BusinessTravel)
Casestudy2NoA$JobLevelFactor = as.factor(Casestudy2NoA$JobLevel)
Casestudy2NoA$WorkYearFactor = cut(Casestudy2NoA$TotalWorkingYears, breaks = c(-1,11,21,40),labels = c("1","2","3"))
Casestudy2NoA$YearsWithCurrManagerFactor  = cut(Casestudy2NoA$YearsWithCurrManager, breaks = c(-1,5,10,17),labels = c("Lo
w","Med","High"))




#Prediction of Attrition for No Attrition Data

Casestudy2NoA$NBPrediction = predict(model,Casestudy2NoA[,c(2,3,19,38,39)])

# Viewing the Prediction

Casestudy2NoA$NBPrediction
```

```
##   [1] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
##  [19] No  No  No  No  No  Yes No  No  No  No  No  No  No  No  No  No  Yes No 
##  [37] No  No  No  No  No  No  No  No  No  No  No  No  No  Yes No  No  No  No 
##  [55] No  No  No  No  No  No  No  No  Yes No  No  No  No  No  Yes No  No  No 
##  [73] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
##  [91] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [109] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [127] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [145] No  No  No  No  No  No  No  No  No  No  No  No  No  No  Yes No  No  No 
## [163] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [181] No  No  No  No  No  No  No  Yes No  No  No  No  No  No  No  No  No  No 
## [199] No  No  No  No  No  No  No  No  No  No  No  No  Yes No  No  No  No  No 
## [217] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [235] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [253] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No 
## [271] No  No  No  No  Yes No  No  No  No  No  No  No  No  No  No  No  No  No 
## [289] No  No  No  No  No  No  No  No  No  No  No  No 
## Levels: No Yes
```

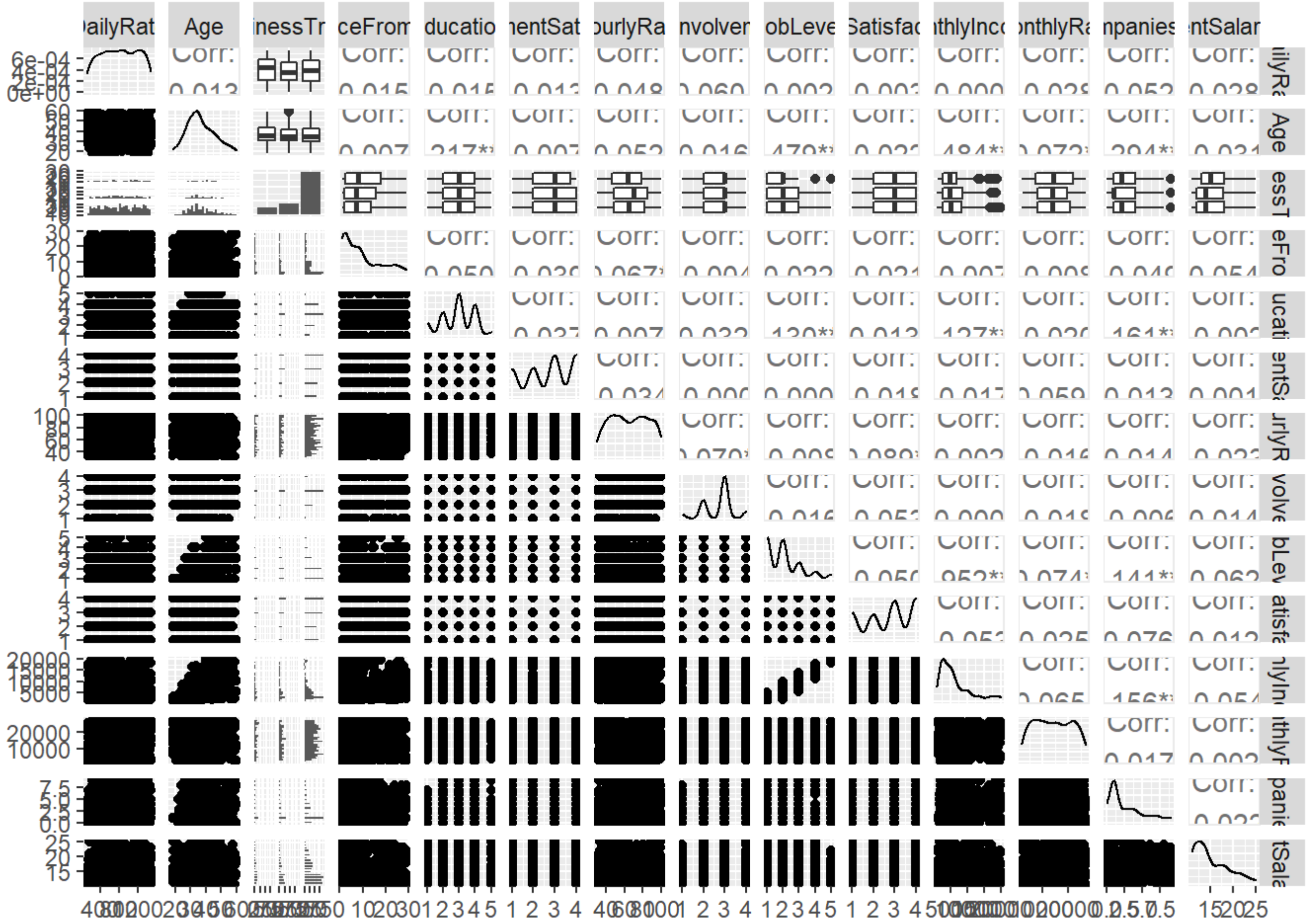
```
#writing csv file for submission
#write.csv(Casestudy2NoA,file = 'C:\\Users\\bhand\\OneDrive\\Desktop\\Doing Data Science\\Case Study 2/Case2PredictionsBhandariAttrition.csv')
```

Salary - Analysis

Bsed on the correlation coefficient, Monthly Income shows evidence of positive relationship with Age (0.484) and Job Level(0.952).

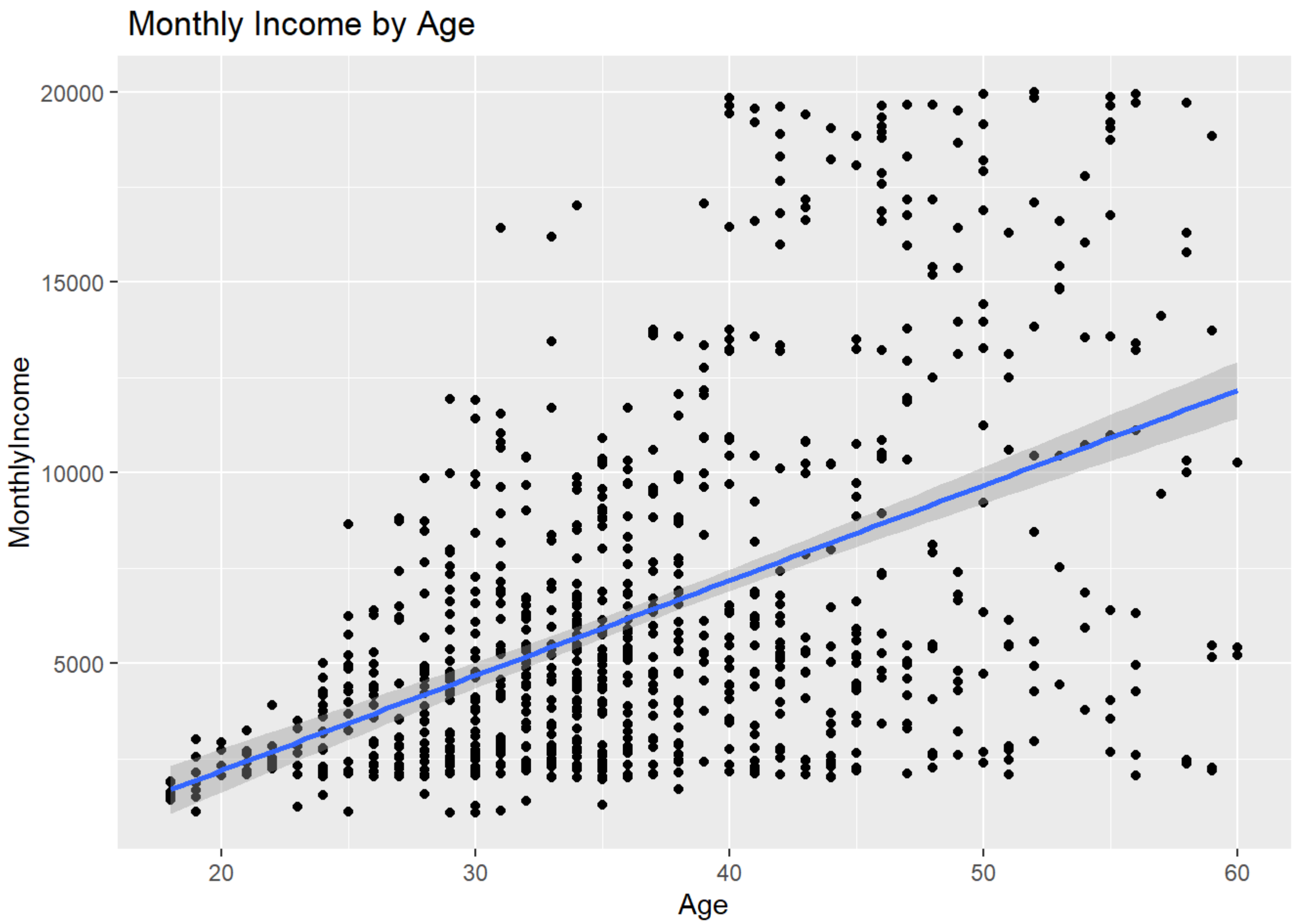
```
data1 <- data
# Selecting quantitative variables
data1 %>%  select(DailyRate,Age,BusinessTravel,DailyRate,DistanceFromHome, Education,EnvironmentSatisfaction,HourlyRate,JobInvolvement,JobLevel,EnvironmentSatisfaction,JobSatisfaction,MonthlyIncome,MonthlyRate,NumCompaniesWorked,PercentSalaryHike)
%>% ggpairs(aes())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



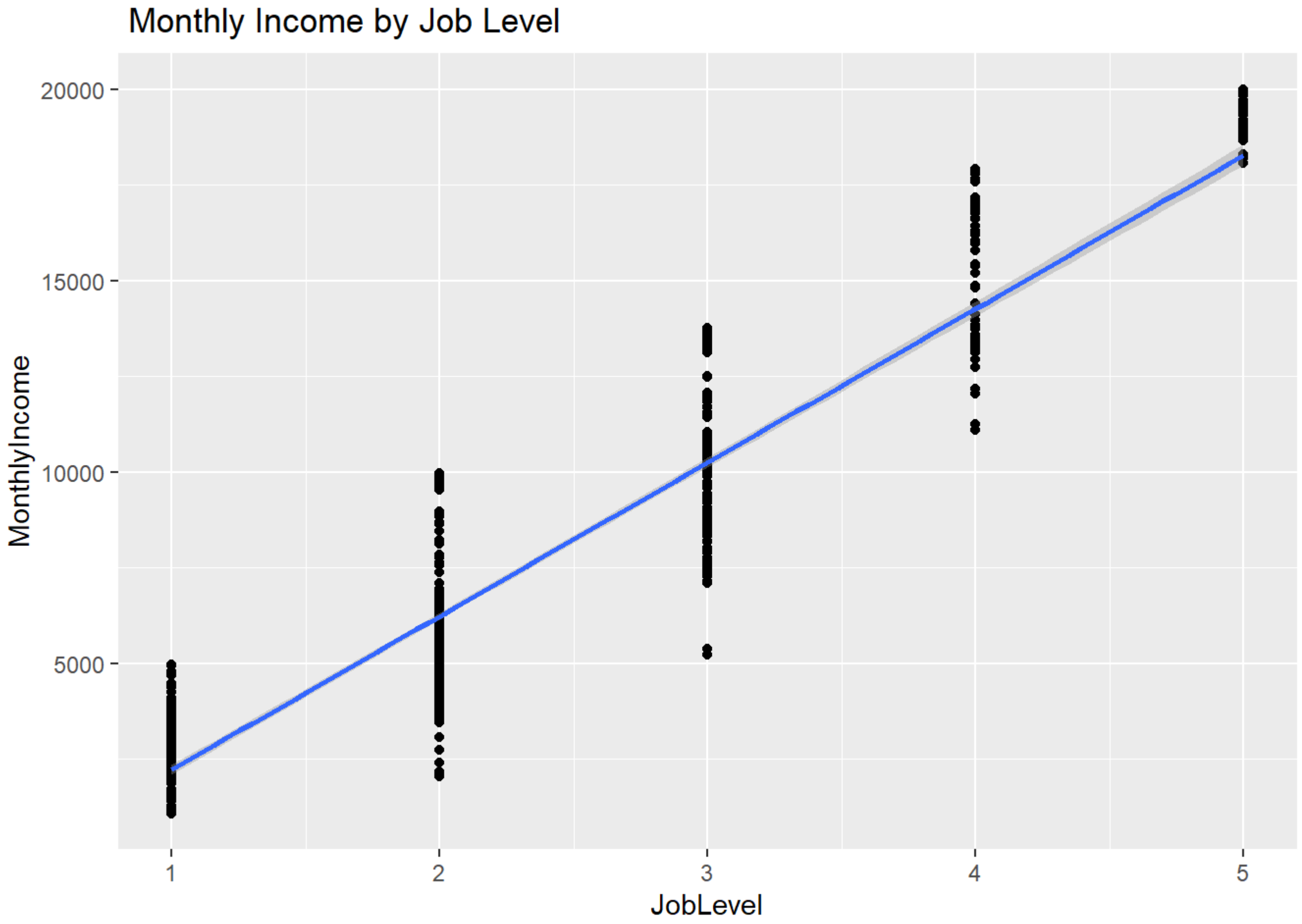
```
data1 %>% ggplot(aes(x= Age, y=MonthlyIncome)) + geom_point() + ggtitle(" Monthly Income by Age") + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
data1 %>% ggplot(aes(x= JobLevel, y=MonthlyIncome)) + geom_point() + ggtitle(" Monthly Income by Job Level") + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Salary - Models

The best linear regression model was model 4 with the lowest RMSE of 1258.839 as well as the best residual density curve.

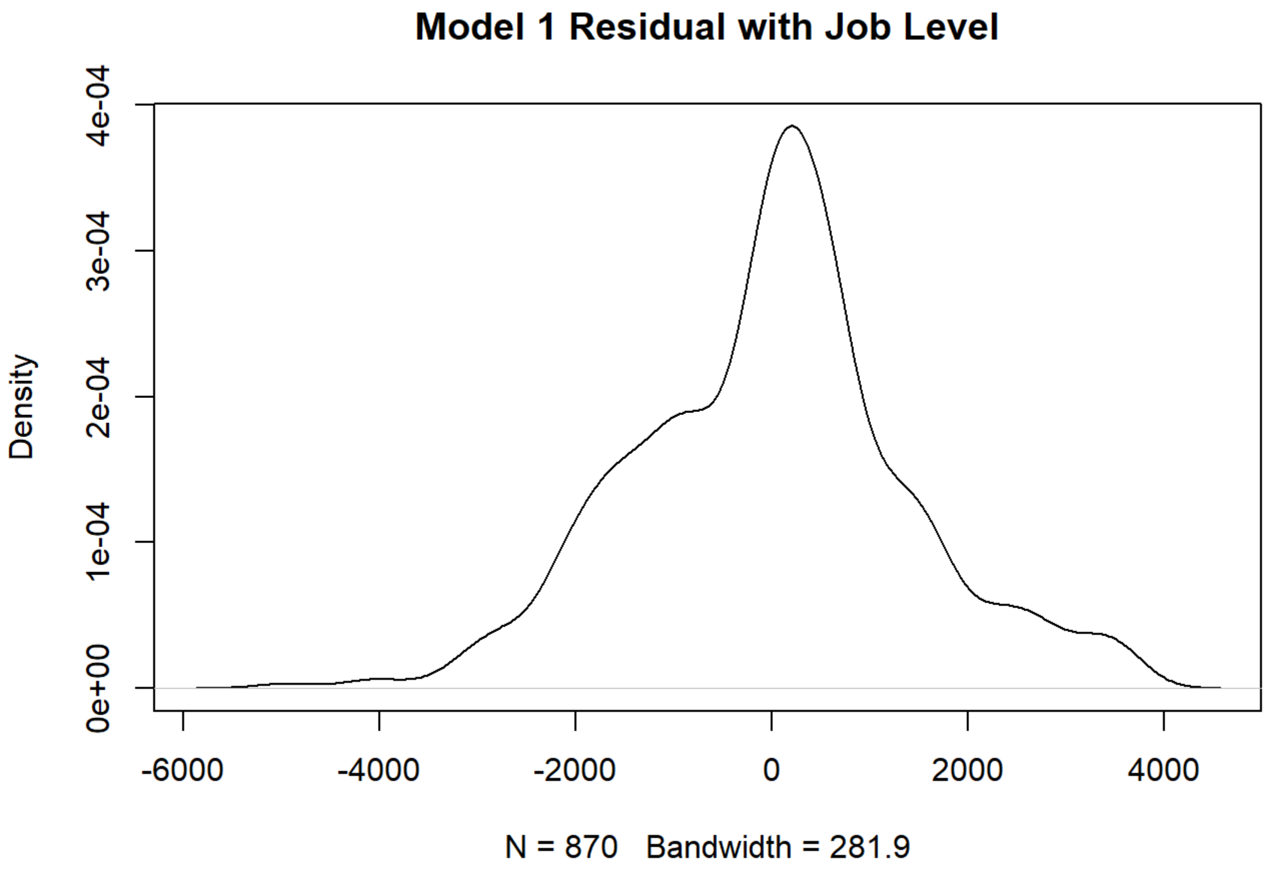
```
# Model 1 with Job level
fit1 = lm(MonthlyIncome~JobLevel, data = data1)
summary(fit1)
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ JobLevel, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5037.1  -928.2    80.1   697.1  3723.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1793.93     101.68  -17.64   <2e-16 ***
## JobLevel     4013.67      43.98   91.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1413 on 868 degrees of freedom
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.9055
## F-statistic:  8329 on 1 and 868 DF,  p-value: < 2.2e-16
```

```
confint(fit1)
```

```
##                 2.5 %    97.5 %
## (Intercept) -1993.494 -1594.375
## JobLevel     3927.352  4099.990
```

```
res1 <-resid(fit1)
plot(density(res1),main = "Model 1 Residual with Job Level")
```



```
RMSE1 = sqrt(mean(fit1$residuals^2))
RMSE1 # 1411.67
```

```
## [1] 1411.67
```

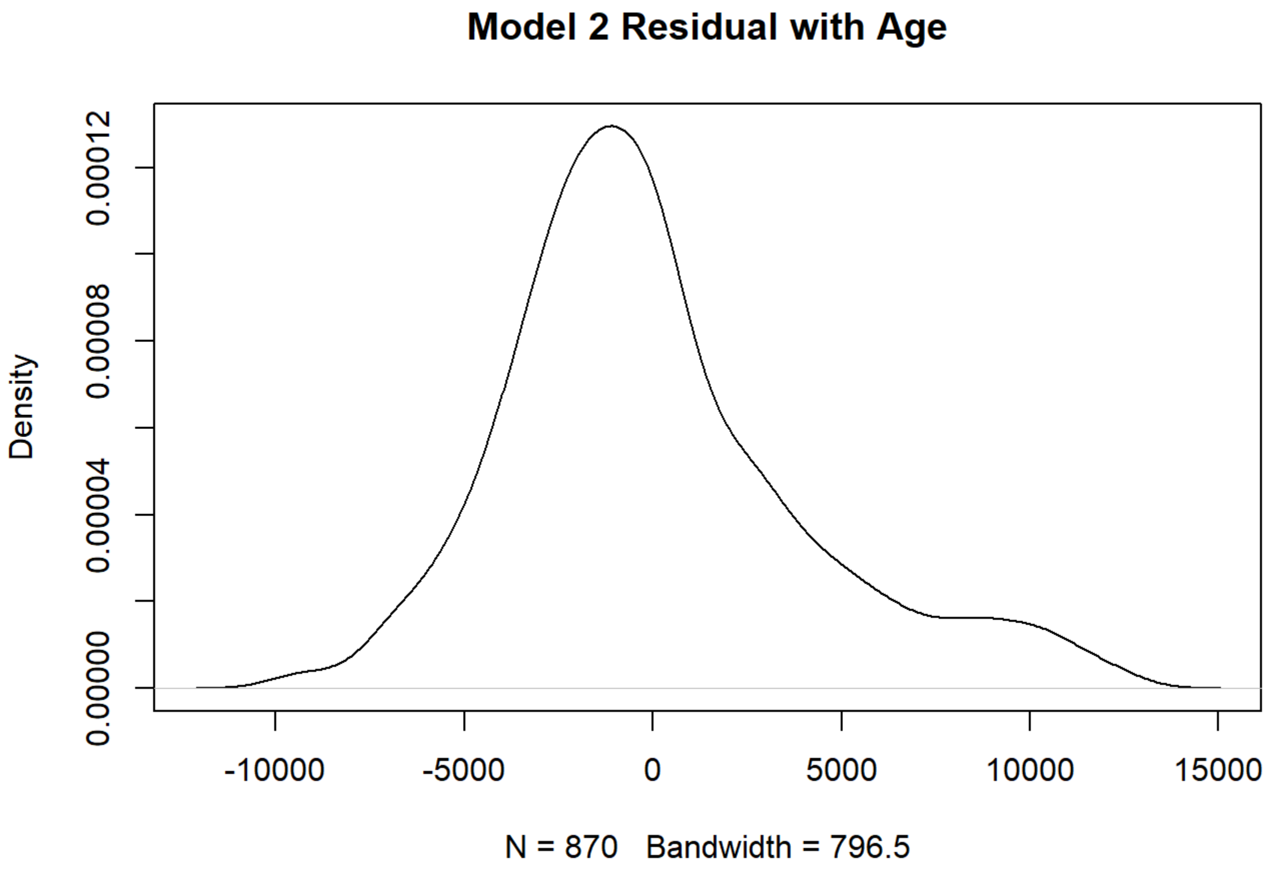
```
# Model 2 with Age
fit2 = lm(MonthlyIncome~Age, data = data1)
summary(fit2)
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ Age, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9744.0 -2622.7  -643.3  1968.7 12651.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2796.8      579.6  -4.825 1.65e-06 ***
## Age            249.4       15.3  16.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4025 on 868 degrees of freedom
## Multiple R-squared:  0.2345, Adjusted R-squared:  0.2337
## F-statistic:   266 on 1 and 868 DF,  p-value: < 2.2e-16
```

```
confint(fit2)
```

```
##                  2.5 %     97.5 %
## (Intercept) -3934.4502 -1659.1417
## Age           219.4314   279.4758
```

```
res2 <-resid(fit2)
plot(density(res2),main = "Model 2 Residual with Age")
```



```
RMSE2 = sqrt(mean(fit2$residuals^2))
RMSE2 # 4020.251
```

```
## [1] 4020.251
```

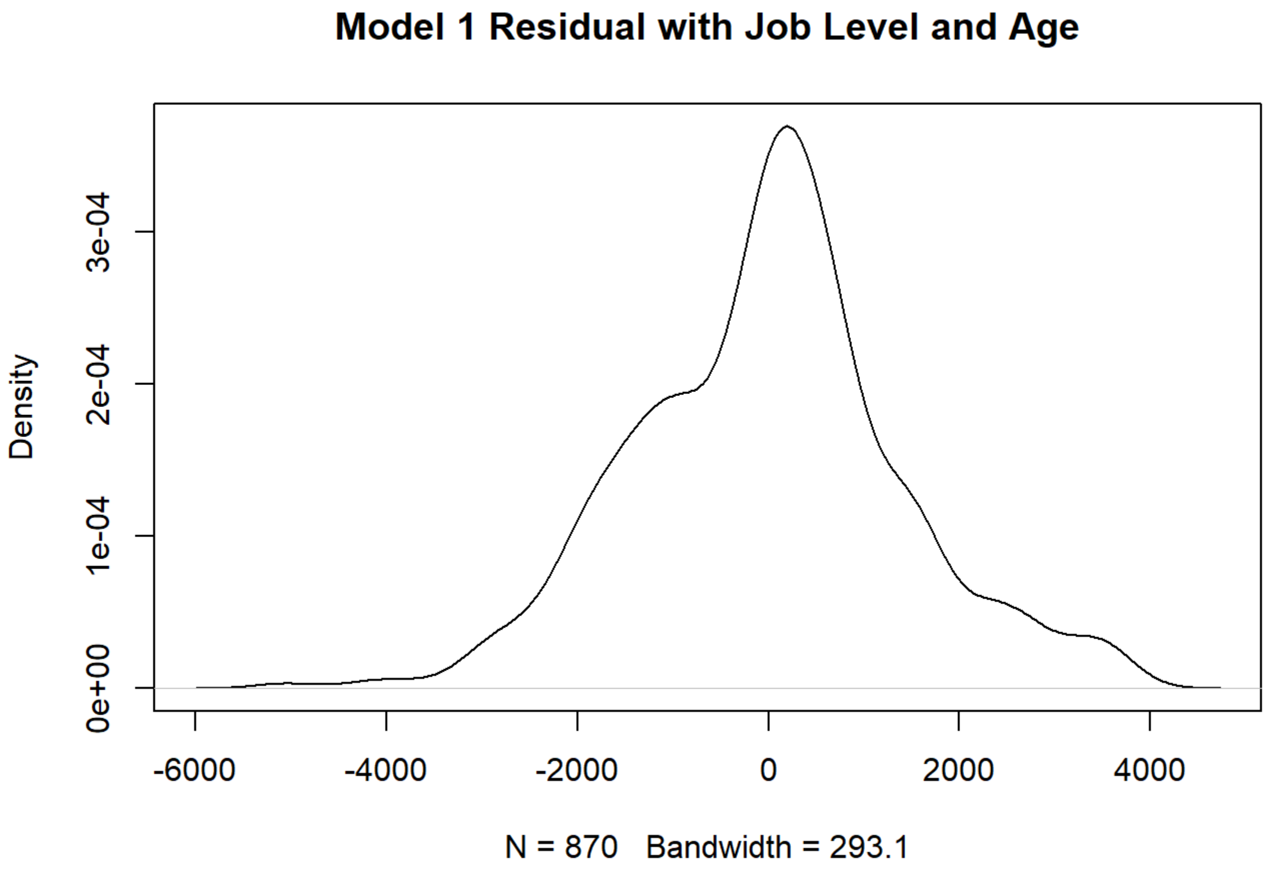
```
# Model 3 combined JobLevel and Age
fitc = lm(MonthlyIncome~JobLevel + Age, data = data1)
summary(fitc)
```

```
## 
## Call:
## lm(formula = MonthlyIncome ~ JobLevel + Age, data = data1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5119.6  -954.7    67.4   734.7  3848.8 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2334.680    202.630  -11.52  < 2e-16 ***
## JobLevel     3940.027     49.871   79.00  < 2e-16 ***
## Age            18.760      6.091    3.08  0.00213 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1406 on 867 degrees of freedom
## Multiple R-squared:  0.9066, Adjusted R-squared:  0.9064 
## F-statistic:  4210 on 2 and 867 DF,  p-value: < 2.2e-16
```

```
confint(fitc)
```

```
##                    2.5 %      97.5 %
## (Intercept) -2732.383300 -1936.97659
## JobLevel     3842.144503  4037.90964
## Age             6.805894    30.71438
```

```
resc <-resid(fitc)
plot(density(resc),main = "Model 1 Residual with Job Level and Age")
```



```
RMSEc = sqrt(mean(fitc$residuals^2))
RMSEc # 1404.01
```

```
## [1] 1404.01
```

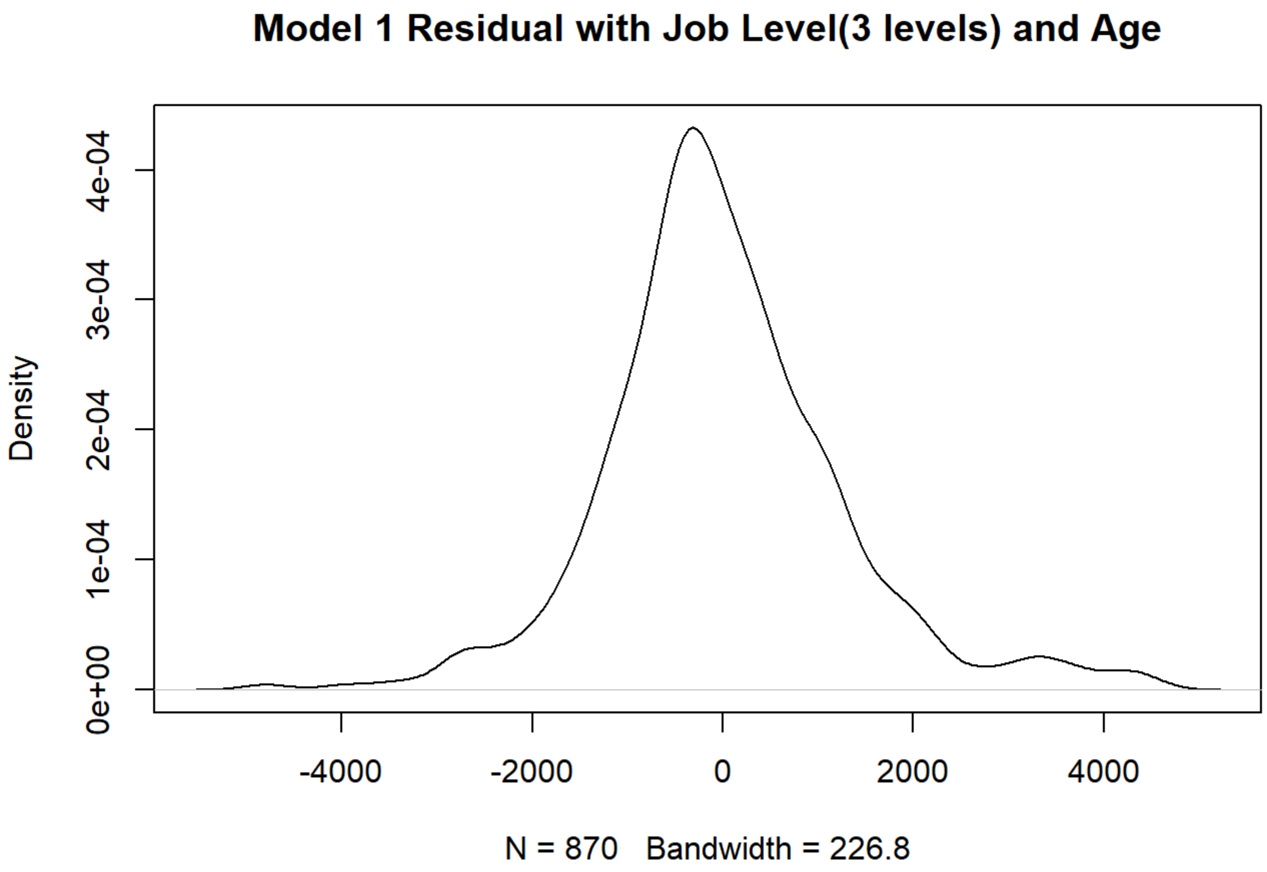
```
# Model 4 combined Job level and Age^2 (Best model with lowets RMSE and highest r squared)
JI2 = data1$JobLevel * data1$JobLevel

JI3 = data1$JobLevel * data1$JobLevel * data1$JobLevel

fitc1 = lm(MonthlyIncome~ JobLevel + JI2 + JI3 + Age, data = data1)
summary(fitc1)
```

```
## 
## Call:
## lm(formula = MonthlyIncome ~ JobLevel + JI2 + JI3 + Age, data = data1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4859.4  -684.7  -121.4   622.5  4542.5 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2546.024    470.355   5.413 8.03e-08 ***
## JobLevel    -2152.035    618.880  -3.477 0.000532 ***
## JI2          2096.029    252.254   8.309 3.70e-16 ***
## JI3          -204.491     30.080  -6.798 1.97e-11 ***
## Age            14.144      5.477   2.582 0.009976 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1262 on 865 degrees of freedom
## Multiple R-squared:  0.9249, Adjusted R-squared:  0.9246 
## F-statistic:  2665 on 4 and 865 DF,  p-value: < 2.2e-16
```

```
confint(fitc1)
```

```
##                     2.5 %     97.5 %
## (Intercept)  1622.854227 3469.19438
## JobLevel    -3366.718176 -937.35251
## JI2          1600.928149 2591.13019
## JI3          -263.530125 -145.45274
## Age             3.393931   24.89355
```

```
res <- resid(fitc1)
plot(density(res), main = "Model 1 Residual with Job Level(3 levels) and Age")
```



```
RMSEc1 = sqrt(mean(fitc1$residuals^2))
RMSEc1 #1258.839
```

```
## [1] 1258.839
```

```
# Adding Predicted Salary based on Model 4 to the No Salary Dataset

Casestudy2NoS$Predicted_Salary_model4<- predict(fitc1, newdata = data.frame(JobLevel= Casestudy2NoS$JobLevel, JI2 =Casestudy2NoS$JobLevel*Casestudy2NoS$JobLevel, JI3 = Casestudy2NoS$JobLevel*Casestudy2NoS$JobLevel*Casestudy2NoS$JobLevel, Age = Casestudy2NoS$Age ), interval = "confidence")

# write.csv(Casestudy2NoS,file = 'C:\\Users\\bhand\\OneDrive\\Desktop\\Doing Data Science\\Case Study 2/Case2PredictionsBhandariSalary.csv')
```

Conclusion

It is extremely difficult to predict attrition. Even though, I was able to create a model with good accuracy, it may not be the best model for prediction, due to various human factors involved. The best way to tackle attrition is to improve the job satisfaction by creating 5 levels for most Job Areas and not just a few.I was able to build a model which can be used to predict salary. This model may be useful to estimate a salary for any new hires based on the current Frito lay data.